Jl - CVI

CENTRAL LIBRARY
26.11.09
UNIVERSITY OF CALCUTTA

THE JOURNAL OF PHILOSOPHY

CUK-H05934-27-KP9800

VOLUME CVI, NUMBER 1
JANUARY 2009

page
5 *Actual Rule Utilitarianism* Richard B. Miller
29 *What Is Lying?* Don Fallis

27

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Michael J. Mooney; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2009

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2009 by the Journal of Philosophy, Inc. ISSN 0022–362X

CEN... L LIBRARY
UNIVERSITY
CALCUTTA
...ORE CAMPUS

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University. It is typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

THE JOURNAL OF
PHILOSOPHY

VOLUME CVI, NUMBER 1
JANUARY–DECEMBER 2009

Edited by
David Albert, Bernard Berofsky, Akeel Bilgrami
John Collins, Arthur C. Danto, Jon Elster
Kent Greenawalt, Patricia Kitcher, Philip Kitcher
Charles Larmore, Isaac Levi, Wolfgang Mann
Frederick Neuhouser, Christopher Peacocke, Carol Rovane
Achille C. Varzi, Katja Vogt

Consulting Editors
James T. Higginbotham, Robert May
Charles D. Parsons, Wilfried Sieg

Managing Editor
John Smylie

Published by The Journal of Philosophy, Inc.

KP2800

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple copying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all copying. Inquiries about copying may be sent to the Managing Editor. Please consult our website, www.journalofphilosophy.org, for more information about our copyright policies.

© Copyright 2009 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and other mailing offices. Please send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. Fax: (212) 932-3721

THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 1, JANUARY 2009

ACTUAL RULE UTILITARIANISM

Some, by no means all, contemporary ethical theorists draw what they think is a useful distinction between ethics and morality. The distinction arguably traces back to Hegel's distinction between *Sittlichkeit* and *Moralitat.* The former is a kind of practical reasoning based on the values and ways of life of actual historical communities. The latter is a more purely philosophical kind of practical reasoning that attempts to transcend the limitations of particular human communities and get to the real rational essence of right and wrong, to principles which hold for all human beings in all times and all places. Among those who draw such a distinction, "ethics" is the name for the rich and complex value judgments grounded in particular communities while "morality" is the name for the thin and abstract judgments of universal right and wrong grounded in pure theory.[1] Influential ethical thinkers such as Bernard Williams and Alasdair MacIntyre have argued for the superiority of ethics over morality. Morality appeals to philosophers who seek universal standards of right and wrong that transcend the limited and local standards of actual communities. But the search for such objectivity, they argue, results in a form of practical reasoning that is dry, abstract, inflexible, and simplistic. Both Kantianism and Utilitarianism are classified as forms of moral theory thus inferior to genuine ethical thinking.

Utilitarianism as it has developed from the work of Bentham and Mill has proceeded as a form of moral theory. The standards of human communities play no essential role in the evaluations given by the Utilitarian. Instead Utilitarian moral theory serves as a standard by which local ethical practices are judged. Differences between the

---

[1] See Robert Piercey, "Not Choosing between Morality and Ethics," *The Philosophical Forum*, XXXII, 1 (2001): 53–72.

0022-362X/09/0601/5–28

ethical judgments of ordinary people and the moral judgments of philosophical theorists, however, tend to cut both ways. While the philosopher may think he is showing the man in the street where he has gone wrong, that stubborn fellow is likely to think that the difference shows instead where the philosopher has gone wrong. Philosophers who adopt the methodology of reflective equilibrium give considerable weight to such ethical judgments, often calling them intuitions. When ethical intuitions clash with moral theory such philosophers see this as a problem for the theory.

From the very beginning Utilitarians have been aware of the tension between such intuitions and the conclusions which follow from their moral theory. Mill in the last and longest chapter of *Utilitarianism* devotes considerable attention to tracing these troublesome intuitions back to their psychological roots in an effort to mitigate their force. In the twentieth century, Richard Brandt and R.M. Hare dismissed the appeal to intuitions as without probative force and proceeded to develop theories without worrying about the supposed demands of reflective equilibrium, broad or narrow. They never intended a moral theory that was limited to describing the ethical standards of their local community but rather boldly proposed a revisionary way of evaluating actions that would be superior to what goes on among the people. As such they clearly preferred morality to ethics.

My own sympathies are with Mill. Revisionary philosophy is, to my mind, vastly more interesting, important, and useful than its timid descriptive cousin. But the criticism that moral theory tends to be dry, abstract, and unrealistic because it is cut off from the practical life of real people is one that I cannot wholly dismiss. MacIntyre's point about the richness and value of tradition should not be ignored. The ways in which communities judge actions are rooted in the experience of many generations. No matter how intelligent or clear-thinking the individual moral theorist, she operates at a distinct disadvantage. Her experience represents nothing more than a tiny and unrepresentative sample of what her community has lived and felt.[2] Every ethical tradition incorporates a vast store of wisdom which the judicious moral theorist would do better to use than to discard.

I do not propose that philosophers use ethical judgments as intuitions to test moral theory in the manner of reflective equilibrium. Nor do I recommend we abandon moral theorizing in favor of expounding the home team's ethical tradition. What I suggest is rather different. I think that the wise moral theorist will incorporate the ethical standards of actual human communities as an essential part of moral theory. Moral theory should not give way to ethics; ethics should be swallowed whole. As the paper proceeds I hope to show how this can be done. At this point a brief description will have to suffice. I think that the wise moral theorist should use actual ethical standards as a basis upon which the more complete, systematic, and useful moral theory is constructed.

---

[2] An oft-divorced colleague of mine once said to me "What do you know of marriage? You've only been married once." It was an odd thing to say but he had a point. No one person's experiences, whether good or bad, match the normal variations encountered in society.

Specifically I will be arguing in favor of a form of Utilitarian moral theory that evaluates actions in terms of their conformity to actual ethical rules. The rules themselves, however, are not accepted uncritically or for the sake of tradition. The concrete ethical standards of real societies should be legitimated by their consequences. The argument will be that such a moral theory represents a pragmatic improvement over actual ethical standards and so the criterion for evaluating moral theory is revisionary and pragmatic and not descriptive or based on intuitions. A brief disclaimer about terminology is also called for. Utilitarianism is normally thought of as a theory which calls for the maximization of utility. I think that optimizing is unwise and that satisficing is far more realistic. So unlike most Utilitarians I am not a maximizer. I considered calling my theory a form of consequentialism but it seems that there is little value in that since the majority of consequentialists these days are maximizers as well. So I will stick with the older term despite its imperfect fit.

What Mill really thought is a matter of controversy with some arguing that he is an Act Utilitarian and others that he is an Ideal Rule Utilitarian.[3] I think he was a Rule Utilitarian who at times came rather close to the position I will argue for here. I certainly think that my argument is one that Mill and Bentham would recognize as kindred to their own.

---

[3] The literature on this point is extensive. See R.B. Brandt, "Some Merits of One Form of Rule-Utilitarianism," *University of Colorado Studies, Series in Philosophy,* III (1967): 39–65, reprinted in Samuel Gorowitz, ed., *Mill: Utilitarianism with Critical Essays* (Indianapolis: Bobbs-Merrill, 1971), pp. 324–44; D.G. Brown, "Mill's Act-Utilitarianism," *The Philosophical Quarterly,* XXIV, 94 (1974): 67–68; Roger Crisp, *Routledge Philosophy Guidebook to Utilitarianism* (New York: Routledge, 1997); Brian Cupples, "A Defense of the Received Interpretation of J.S. Mill," *Australasian Journal of Philosophy,* L, 2, (1972): 131–37; Alan E. Fuchs, "Mill's Theory of Morally Correct Action," in Henry R. West, ed., *The Blackwell Guide to Mill's Utilitarianism* (New York: Oxford, 2006), pp. 139–58; Alasdair MacIntyre, *A Short History of Ethics: A History of Moral Philosophy from the Homeric Age to the Twentieth Century* (Notre Dame: University Press, 1998); J.O. Urmson, "The Interpretation of the Moral Philosophy of J.S. Mill," *The Philosophical Quarterly,* III, 10 (1953): 33–39; Alan Ryan, "Introduction," in Mill and Bentham, *Utilitarianism and Other Essays,* Ryan, ed. (London: Penguin, 1987), pp. 7–63; William Shaw, *Contemporary Ethics: Taking Account of Utilitarianism* (New York: Oxford, 1999), p. 165.

The revisionary way to argue for the superiority of Rule Utilitarianism is to emphasize the utility of rules as such and to stress how violations of rules, even for the sake of predictable consequences, undermine the rules. Rules can only tolerate so much flexibility or they completely break down. Act Utilitarianism is incompatible with the degree of rigidity that rules require. The truth of the matter is paradoxical. To aim directly at human flourishing is to miss the mark.[4] The goal can be approximated only by the indirect route of conforming to rules that promote the social order. Hence there is an air of paradox about Rule Utilitarianism that makes it seem like "rule worship," but no real problem.

If the Utilitarian is convinced of the need to evaluate actions according to their conformity to rules and judges the moral authority of rules according to their utility, an important question remains: Are the rules to which the Rule Utilitarian refers the set of merely possible rules which would maximize utility if human societies adopted them, or those actual moral rules of actual societies which at least promote the general welfare? The Ideal Rule Utilitarian gives the first answer; the Actual Rule Utilitarian gives the second. Since Rule Utilitarianism is almost exclusively thought of in the former way, a brief explanation of how Ideal Rule Utilitarianism came to be the preferred form is in order.

In 1936 R.F. Harrod argued for a precursor of Actual Rule Utilitarianism.[5] His arguments emphasize the need for public understanding and enforcement of practices and that the value of these practices typically depends on actual general compliance. In 1955 John Rawls's "Two Concepts of Rules" once more refocused attention on the importance of rules, practices, and institutions in utilitarian thought.[6] The time seemed to be ripe for a fresh appreciation of how the greatest happiness principle could be employed as the criterion to evaluate the actual moral rules of a society which in turn are the criterion for the morality of individual action. However, when J.J.C. Smart in "Extreme and Restricted Utilitarianism" (1956) challenged Rule Utilitarianism, he focused on Ideal Rule Utilitarianism.[7] His argument that it collapsed into Act Utilitarianism did not convince everyone, but it gave the impression that Rule Utilitarianism was identical with Ideal Rule Utilitarianism. David Lyons's *Forms and Limits of Utilitarianism* (1965) discussed in detail many versions of Ideal Rule Utilitarianism and so helped to solidify this impression.[8] D.H. Hodgson devoted a section of *Consequences of Utilitarianism* (1967) to a critique of Actual Rule Utilitarianism. Hodgson argues that it cannot be a complete system of ethics and that the gaps must be filled either in a totally ad hoc way or by reverting to Act Utilitarianism.[9] After Hodgson, Actual Rule Utilitarianism was largely forgotten. There are some small indications, however, that the time is ripe for a fresh reconsideration of Actual Rule Utilitarianism. Brandt, perhaps the most distinguished defender of Ideal Rule Utilitarianism, in the final formulations of his position moved in the direction of Actual Rule Utilitarianism. Brandt admitted that the ideal rules must often mention the actual rules of the agent's society as relevant conditions to be considered. Another Ideal Rule Utilitarian, Brad Hooker, reluctantly allows that we should compromise with actual ethical conventions when doing so would prevent disastrous amounts of harm to innocent persons.[10] But the closest approach to Actual Rule Utilitarianism in recent times is the position of Conrad Johnson in his fine book *Moral Legislation*.[11] Johnson gives powerful arguments in favor of the relevance of actual ethical rules. Later in this paper I will borrow these arguments from Johnson while criticizing him for not going far enough in this direction.

---

[4] The paradox here parallels one in epistemology: to maximize truth we need to aim at justified beliefs and not directly at true beliefs.

[5] Harrod, "Utilitarianism Revised," *Mind*, XLV (1936), reprinted in Gorovitz, pp. 73–87.

[6] Rawls, "Two Concepts of Rules," in Gorovitz, pp. 175–94.

[7] Smart, "Extreme and Restricted Utilitarianism," in Gorovitz, pp. 195–203.

Despite this, it is undoubtedly the case that the most popular form of Rule Utilitarianism at present is Ideal Rule Utilitarianism. Ideal Rule Utilitarianism judges actions not directly by their consequences but indirectly by their consistency with the ethical rules which it would be ideal for society to adopt. The most familiar objection to Ideal Rule Utilitarianism is that, without some constraints on what rules to consider as possibly ideal, it collapses to Act Utilitarianism. For without any constraints the rules may grow indefinitely complex, taking in more and more of the details that are relevant to the consequences of the action until each action has its own rule. Brandt has devoted great care in crafting a version of Ideal Rule Utilitarianism that blocks the collapse to Act Utilitarianism.

---

[8] Lyons, *The Forms and Limits of Utilitarianism* (New York: Oxford, 1965).

[9] Hodgson, *Consequences of Utilitarianism* (New York: Oxford, 1967), pp. 26–32.

[10] Hooker, "Compromising with Convention," *American Philosophical Quarterly*, XXXI, 4 (1994): 311–17. Hooker has been quite active recently in promoting Rule Utilitarianism and defending it from objections. His other recent writings include: "Rule-consequentialism, Incoherence, Fairness," *Proceedings of the Aristotelean Society*, XCV (1995): 9–35, "Is Rule-Consequentialism a Rubber Duck?" *Analysis*, LIV, 2 (1994): 92–97, "Rule-Consequentionalism and Demandingness: A Reply to Carson," *Mind*, C (1991): 269–76, and "Right, Wrong, and Rule-Consequentialism," in West, ed., *The Blackwell Guide to Mill's Utilitarianism*, pp. 233–48.

[11] Johnson, *Moral Legislation* (New York: Cambridge, 1991).

Rule Utilitarianism can also be attacked on grounds that the widespread adoption of Rule Utilitarianism would have bad consequences. It is objections of this sort that a pragmatic revisionist will take most seriously. Brandt is able to stop the slide into Act Utilitarianism, but his appeal to ideal rules fails to establish a system which would actually be good for people. Failing to conform to ethical intuitions is nothing to worry about; however, if utilitarianism fails to produce good consequences it is an unacceptable proposal.

Brandt understands the consequentialist objections to various forms of utilitarianism. In fact his statement of some of these objections could scarcely be improved: "[Act Utilitarian] calculations are so subtle as to constitute a standing invitation to rationalize in one's own favor .... As a result there would not be the security in predicting how others would behave necessary for making personal plans or effectuating schemes of cooperation" (*op. cit.*, p. 142). That there are important goods which can be reliably achieved only through the existence of rules governing the behavior of individuals has been widely recognized in discussions of utilitarianism for at least thirty years.[12] But the point goes back much farther, at least to Hume's striking analogy of two men pulling oars in a common boat.[13] There are certain goods that depend on coordinated action, which are necessarily lost when individuals act according to their own judgment.

For example, consider cases in which the public good requires private sacrifice by many individuals. Suppose some important public good could be achieved by having every private home turn down its thermostat to 60 degrees. If individuals decide for themselves how far back to reduce their level of comfort, it is unlikely that the public good will be served. Thus we need a rule to tell individuals how much sacrifice is required. Moreover, the publicity of the standard helps persuade each individual that others are in fact making the needed sacrifice. Without rules, individuals will have to use their own judgment to determine how much sacrifice is appropriate. Individual judgment unfettered by public discussion is helpless against rationalization. Moreover, even if the agent is fully aware of the natural tendency to rationalize and successfully guards against it in her own case, this very fact makes her more likely to believe that others will delude themselves with self-serving rationalizations. So, left to their own discretion, neither the naïve nor the sophisticated will see reason to make the sacrifices necessary to secure the public good. In summary, "collective supervision and guidance of behavior is different from self-supervision and guidance."[14] The former is often preferable and sometimes necessary. Hence people need rules.

---

[12] See Lyons, *The Forms and Limits of Utilitarianism*, p. 140, and Bernard Williams, "A Critique of Utilitarianism," in Smart and Williams, eds., *Utilitarianism: For and Against* (New York: Cambridge, 1973), pp. 77–150.

[13] Hume, *A Treatise of Human Nature*, Book III (New York: Oxford, 1973), p. 490.

Brandt's appeal to ideal rules is not sufficient to solve the problems of rationalization and coordination. First, consider the rationalization problem. Both the probability and the seriousness of possible injuries to others are consistently underestimated. But if the individual is asked to conform his behavior to an ideal system of rules, there is no reason to think that the agent's personal concept of this will be any less pliant than the agent's personal hedonic calculus. Second, consider the coordination problem. In a society of Act Utilitarians I will often be unable to determine how others will behave because I will be unable to predict the results of their utilitarian calculations. It will be practically impossible for me to guess each agent's background knowledge and biases. Act Utilitarian calculations sensitively depend on subjective factors and this precludes coordinated action. The appeal to ideal rules does not significantly change this. A society of Ideal Rule Utilitarians would possess such various and subjective concepts of the ideal rules that coordination would be just as impossible. Only actual public rules provide the intersubjective stability necessary to solve the rationalization and coordination problems. Thus two important reasons which Brandt has for preferring Rule Utilitarianism to Act Utilitarianism turn out upon careful consideration to support Actual Rule Utilitarianism rather than Ideal Rule Utilitarianism.

Another way to see this is to study Brandt's attempts to solve these problems in the context of ideal rules. The following examples illustrate the problem for ideal rules. Suppose that the system of rules which would be ideal for human society is one in which bicycles were a public utility rather than private property. If right actions are actions which conform to the ideal rules then I should act as though bicycles are already a public utility and not private property. But it is plain that if I acted in this way much harm and little good would result. A similar but even more extreme example is found in the conventions of driving. In America we drive on the right-hand side of the road. In Britain they drive on the left. But the majority of drivers are right-handed which means that it is probable that one of these rules is superior in practice to the other. Suppose it were true that, all other things being equal, a system of driving on the left is safer. If I were obliged to act according to the ideal system of rules that would mean that I should drive home on the left no matter where I live. This is surely not a welcome conclusion for a philosophy which values human happiness.

---

[14] Johnson, p. 33.

The problem forces Brandt to make a place for actual rules within the context of his ideal rules. "An optimal morality will take this structure [of actual institutions with their recognized rights and duties] for granted in its provisions for everyday behavior" (*op. cit.*, p. 149). In other words Brandt's ideal rules will mention the actual rules of the society and with respect to most actual rules, recommend normal obedience even if the rules are themselves less than optimal. So he blocks the absurd result that we should act now as if bicycles were a public utility rather than private property by modifying ideal rule utilitarianism in the direction of Actual Rule Utilitarianism. Similarly, the ideal rules presumably would also recommend that we conform our driving behavior to the actual rules of the society in which we live rather than to the ideal driving rules. Thus Brandt's ideal rule utilitarian makes right actions depend on actual rules and not merely on possible ideal rules. It is the actual rules that are producing the good effects of coordinating behavior and providing a hedge against rationalization. Ideal rules are useless theoretical appendages. In the real world they are actually harmful.

Similarly, Hooker's attempt to make a reasonable compromise with actual ethical rules fails.[15] Hooker notes that any viable ethical system of rules will have a rule about preventing disasters. He concludes from this that the sensible Ideal Rule Utilitarian will go along with imperfect actual rules when violating them would have terrible consequences. This is unobjectionable but fails to come to terms with the full force of the objection. Hooker's response will work for extreme cases like the convention about the side of the road on which to drive, for the consequences of flaunting this convention are indeed extreme. However, this is not the case for many conventions, including the one which treats bicycles as private property. If I started unilaterally acting as if bicycles were a public utility, this would inconvenience people who believed that they owned bicycles. But it would be a gross exaggeration to call such inconvenience a "disaster." And if the Ideal Rule Utilitarian were tempted to stretch the rule about preventing disasters to cover such minor consequences, then her position would have almost completely collapsed back into Act Utilitarianism again.

We turn now to the consideration of a moral theory which makes a stronger and more consistent appeal to actual rules than do Brandt or

---

[15] Hooker, "Compromising with Convention."

Hooker. Johnson conceives of morality on the model of the law. I find much to praise in Johnson's theory. His explanations of the importance for morality of actual rules publicly known and collectively enforced can hardly be improved upon. He begins by noting that the value of widespread acceptance of a rule is generally greater than the value of equally widespread behavior consistent with a rule. There is a surplus causal effect of rule acceptance. Rules help people to achieve coordinated behavior, but that is only half the story. The other half is the value of trust. "The good of mutual trust may fail to exist even when everyone's behavior is in outwardly strict conformity" (*op. cit.*, p. 28). For example, when conformity is due to the fear of punishment by external authorities rather than from the general voluntary acceptance of the standards, something important is lost. People value the sense of community which is comprised by such standards and which should not be confused with coerced conformity. This community is reinforced by shared commitment to common rules, which are maintained by tying moral praise and blame to those rules.

So rules help ethical communities to minimize rationalization, to gain the values of coordinated effort, and to build a sense of community and trust. Nor is this all. Johnson adds that reliance on rules promotes freedom and security for individuals. "For it is important to people, living in community with others, to know what they need to do in order to be in good standing with others, and to be confident that they have done their duty and that they therefore can rightly claim that they are not to be blamed or criticized" (*op. cit.*, p. 31). Johnson's comparison of morality to law is germane here. The rule of law protects individual freedom by assuring each person that they will not be prosecuted for anything other than a violation of the law and that the laws are clear enough so that no one needs to be puzzled about whether or not his actions violate the law. Laws must be public to promote coordination and community while preserving liberty and security. "What is often needed, therefore, is something approximating a collective understanding, first about what is to count as morally wrong behavior, and secondly that blame, when otherwise appropriate, will be directed toward people for the good that a particular instance of blame might do" (*op. cit.*, p. 32).

While Johnson's position recognizes the importance of actual rules it does not go far enough. Just as Brandt's latest system can be understood as a modification of Ideal Rule Utilitarianism that makes room for actual rules, Johnson's system can be understood as a form of Act Consequentialism that makes room for the real ethical conventions of society. Johnson's system differs from Actual Rule Utilitarianism in three ways, each of which weakens his system. First, Johnson bases moral judgment fundamentally on a version of Act Consequentialism which he calls the primary principle. The primary principle is "promote good and avoid evil." It is meant to be less definite than classical Act Utilitarianism which requires agents to maximize the good. Rules enter the picture for Johnson only as constraining the primary principle; that is, rules tell agents when they should not conform their actions to the primary principle itself. Johnson's system would seem to confirm Hodgson's prediction that the Actual Rule Utilitarian will fall back on Act Utilitarianism to fill in the gaps between the rules. The second departure from Actual Rule Utilitarianism lies in his appeal to the primary principle as the background against which rules should be assessed. "A rule R is justified for G if the general acceptance of R in G produces greater overall good than that which would be produced if members of G governed their actions in that sphere by use of the primary principle, along with appropriate individual strategies, and without use of any collective rule like R" (*op. cit.*, p. 38). The third area in which Johnson departs from Actual Rule Utilitarianism lies in his claim that "we must also recognize that large areas of our conduct and moral thinking are not suited to governance by rules. Or, if rules are appropriate in these areas, they are not the sort of rules that are understood to be applicable to everyone. Rather, they are rules that are understood to be largely a matter of individual choice and creation" (*op. cit.*, p. 83). The strength of Actual Rule Utilitarianism can be made clearer if I criticize each of Johnson's three departures from it.

First I object to Johnson's treating the actual rules merely as modifications and restrictions on Act Consequentialism—his "primary principle." I think that Johnson ought not to have fallen for Hodgson's gambit. The greatest happiness principle is not a rule but a criterion for evaluating rules. Instead of seeking a "complete" system which will tell agents what they must do in every situation, a reasonable ethics will forbid certain classes of action while at the same time creating a generous sphere of permissible actions in which agents can do as they please. The gaps between the rules are the spaces in which responsible persons are free. Freedom is not a defect needing remedy. Johnson should note that the security from blame which he prizes is lost if his primary principle is allowed to function as a rule. The primary principle is a vague law of wide applicability. If subject to such a law could I generally be confident that I am free from blame? On the contrary, because of the subjectivity of such calculations my actions would frequently be second-guessed by others invoking the primary principle. It may be true that if specific ethical rules are considered as limitations on the primary principle, then insofar as those specific ethical rules applied I might have the confidence of which Johnson speaks.

But in all the areas in which no specific rules supersede the primary principle, insecurity prevails. Moral theory should not seek to prescribe a single unique right action for every situation. Moral theory should not seek to micromanage the lives of responsible people. Members of the ethical community should not be condemned or criticized for actions that are consistent with the ethical rules of their community. In the absence of any rule covering the situation, agents are at liberty to do as they please. Johnson's use of the primary principle to fill the gaps in actual ethical systems is a solution in search of a problem.

Johnson also uses the primary principle to explain how consequences can be used to evaluate the ethical rules of society. Any utilitarian appeal to consequences is necessarily comparative. So one question any utilitarian must answer is: What is the baseline against which the comparison is being made? Johnson's baseline is the primary principle. Actual ethical rules are measured against what would happen if people were to govern their actions by use of the primary principle instead.[16] Johnson's assumption that people will conform to the primary principle in the absence of specific rules is both unrealistic and unwise. It is unwise because the direct use of the greatest happiness principle by individuals is often harmful. It is unrealistic in that the standard it requires is too high. Realistically we do not expect normal people to think much about what is in the public interest. People spend their lives pursuing the good life for themselves and their children. If they are ethical, they do not lie, cheat, or steal in pursuit of those ends. But personal goals are the warp and woof of people's lives. Common sense dictates that what is not forbidden is permitted. Suppose that my ethical community has adopted a rule whose consequences are of dubious value. In deciding whether elimination of the rule would be an improvement we must realistically predict what normally conscientious people would do in the absence of the rule. Realistically we ought to expect that in the situations formerly covered by the rule individuals would not revert to act utilitarianism but rather would pursue self-interest or the interests of friends and family. The ethically conscientious person pursues self-interest within the limits drawn by her society's ethical conventions. When one ethical rule breaks down the scope for self-interested action simply expands, constrained only by the remaining ethical conventions.

The third point on which I differ with Johnson is relatively minor. Johnson believes "that large areas of our conduct and moral thinking are not suited to governance of rules" (*op. cit.*, p. 83). I agree that there are areas of our life intimately concerned with values where the appropriateness of general rules is questionable. Johnson's two examples of individualistic goods are the responsibility to develop one's talents and the responsibility to give something to charity. People's talents and opportunities differ markedly. Consequently what will be effective varies greatly from individual to individual. The rules which would help individuals to develop their talents are as various as the individuals themselves. Similarly, to offer one's time, money, or energy to help the needy is an important social good. But the details of when and to whom and how much to give are best left to the individual. Any specific rule, like tithing, is too specific. There is good reason to think that these imperfect duties should be left unspecific.

---

[16] This is an abbreviated version of Johnson's position quoted more fully on page 13.

Yet I cannot agree with Johnson that social rules are completely out of place here. Coordination is not the only good that general social rules promote. As Johnson himself emphasizes, publicly acknowledged rules promote a sense of community which is a surplus causal effect of rules. The general rules covering the development of talent and requirements of charity must be very general; however, very general rules are rules nonetheless. Such collective rules have a positive role to play even in areas of life in which it is wise to grant to the individual wide discretion. Johnson writes convincingly of the value of membership in an ethical community. "Collective supervision and guidance of behavior is different from self-supervision and guidance" (*op. cit.*, p. 33). The benefits of the former are not limited to the informal punishments that the ethical community can impose on free riders to discourage them from doing wrong. The benefits include the encouragement and support which the community can offer to help individuals to do positively what is right. The public recognition of highly general rules about self-development and charity helps individuals feel a sense of solidarity with their ethical community and serves as the reminder that well-intentioned but weak-willed ordinary human beings often need. Public rules promote a higher average level of effort with correspondingly better effects than would be the case if the matter were left purely to individual discretion.

To summarize so far, I have shown that there are four reasons why revisionary moral theory implies the superiority of a system based indirectly on rules as opposed to one based directly on individual actions: (1) rules provide a hedge against rationalization, (2) some important goods can only be achieved through coordinated effort, (3) collective rules create a sense of belonging to a moral community, and (4) explicit rules secure individuals from moral blame. I have shown how all four advantages are in reality advantages of actual rules and not advantages of merely possible rules. Merely possible rules, no matter how ideal, can never create more than possible trust, possible coordination, possible community, or possible security. I have further argued that there are no good reasons why the appeal to actual rules should operate merely as a limitation on Act Consequentialism. Rather I have shown how the ethical authority of actual rules which promote human flourishing is an essential part of moral theory. From my perspective Ideal Rule Utilitarianism is too utopian. That is to say it forgoes the limited but real value of actual rules in favor of the dubious value of supposedly more perfect but imaginary rules.

It is beyond the scope of this paper to criticize current versions of Act Utilitarianism; nevertheless, I will indicate why Act Utilitarianism is not a live option for revisionary ethics. The Act Utilitarian's most popular criticism of Rule Utilitarianism is Smart's "rule worship" objection, which originally appeared in 1956 and continues to resurface even today. The continued popularity of this objection is hard to explain in the light of decades of discussion of the utility of rules. The appeal to rules is not unmotivated or motivated only by a desire not to run afoul of ethical intuitions. The motivation is hard-headedly pragmatic: rules are necessary for human flourishing. Act Utilitarianism is simple and direct but destructive of practices which are essential to the welfare of human societies. Considered as a revisionary moral theory, it is a disaster. This argument, which I will call the Consequence Argument, was made by Hodgson in *The Consequences of Utilitarianism* (*op. cit.*). Act Utilitarianism is an unacceptable moral theory because it fails to deliver the goods.

One possible Act Utilitarian response to the Consequences Argument is to appeal to expectations of mutual cooperation of other beneficent agents. One should not simply do whatever one judges will produce the best consequences, but rather one should do whatever would produce the best consequences if adopted as part of a scheme of cooperation by others of good will. I think such a modification of Act Utilitarianism really amounts to admitting the superiority of Rule Utilitarianism. Such a theory does not evaluate actions in terms of the particular consequences of particular actions, which is the essence of Act Utilitarianism, but rather evaluates actions according to the consequences of a system, which is the essence of Rule Utilitarianism. Such theories bifurcate, just as Rule Utilitarianism does, into two forms depending on whether the expectations are about ideal but merely possible systems of cooperation[17] or whether they are what we expect morally decent people to do given the actual moral conventions of their society. If the former, this theory is a form of Ideal Rule Utilitarianism. If the latter, the theory is Actual Rule Utilitarianism. So nothing more needs to be said about expectation-based Act Utilitarianism.

---

[17] For a system of this type see Donald Regan's *Utilitarianism and Cooperation* (New York: Oxford, 1980).

At this point I will turn to two related tasks: (1) a precise formulation of Actual Rule Utilitarianism and (2) the attempt to anticipate and answer plausible objections. Brandt and Fred Feldman argue that Actual Rule Utilitarianism fails because the de facto rules of human societies are clearly unacceptable to enlightened common sense.[18] Hodgson in an admirable but totally forgotten discussion of Actual Rule Utilitarianism offers two more objections: (1) it is too conservative insofar as it permits but does not require individuals to actively prod their society into improving its moral conventions (*op. cit.*, p. 29–30), (2) it is incomplete insofar as conventions may be indefinite, conflicting, or prescribe only minimal standards of behavior (*op. cit.*, p. 30). I shall answer these objections in the process of giving a more precise statement of Actual Rule Utilitarianism.

A convenient starting point is Feldman's formulation of the position. Feldman understands the position as follows:

> (ARU1) An act, *a*, is morally right if and only if there is some practice *P*, such that *P* governs the situation in which a would occur; *P* permits *a*; and *P* is de facto in the society in which *a* would be performed (*op. cit.*, p. 76).

Feldman refutes ARU1 as follows: "Surely, we do not want to say that a man who owns and mistreats slaves acts rightly simply because his fellow citizens accept the practice of slavery! Their belief in slavery surely does not serve to raise the normative status of his act .... In general we must conclude that [Actual Rule Utilitarianism] is not a form of utilitarianism at all. It is merely a form of *conventionalism*" (*op. cit.*, p. 76).

Actual Rule Utilitarianism, like conventionalism, appeals to the actual rules of the agent's society. Is it not then merely a form of conventionalism subject to the same devastating objections? Parenthetically, I note that the revisionist understands the objection to conventionalism not as its clash with intuitions about the propriety of slavery, but rather as its inability to be selectively critical of social conventions. The conventionalist must accept both harmful and valuable social institutions on the same footing. Hodgson noted that the Actual Rule Utilitarian has an easy and obvious answer to this objection. The Actual Rule

---

[18] Brandt, in Gorovitz, pp. 324–44; Fred Feldman, *Introductory Ethics* (Englewood Cliffs, NJ: Prentice-Hall, 1978).

Utilitarian, unlike the conventionalist, accepts the ethical rules of a society only when those rules produce good results (*op. cit.*, p. 27). Actual Rule Utilitarianism does not uncritically accept all actual ethical rules. Hodgson noted "It seems most plausible to require that a conventional moral rule be justified by its utility in the sense that the consequences of its break-down or disappearance (or replacement by a *permissive* rule) should be worse than those of its maintenance" (*op. cit.*, p. 27). It is reasonable for Feldman to point out the similarity of Actual Rule Utilitarianism to conventionalism, but fairness also requires that we not forget that the principle of utility is being employed to evaluate the rules as is the case with any form of Rule Utilitarianism.

A technical problem with Feldman's definition is that it makes actions right only if there is a rule governing the situation in question. Often there is no such rule. For example, there is no moral rule about which shoe to put on first. The consequence that neither way of putting on my shoes is right is awkward. What is wanted is a definition that makes both ways right. A better definition would require that there be no rule forbidding the action rather than requiring that there be a rule permitting it.

Our second formulation takes these points into consideration.

(ARU2) An act, *a*, is morally right if and only if it is not the case that there is a practice, *P*, such that *P* governs the situation in which *a* would occur; *P* forbids *a*; *P* is de facto in the society in which *a* would be performed, and *P* promotes happiness.

Hodgson's second objection was that the system was necessarily incomplete. In my earlier discussion of Johnson, I answered a simplified version of this objection. Since Hodgson develops the objection in more detail, I will expand on that answer here. Hodgson finds Actual Rule Utilitarianism incomplete due to three problems: (1) indeterminacy in the rules themselves, (2) possible conflicts among the rules, (3) conventional moral rules requiring only "the minimum necessary for satisfactory communal existence" (*op. cit.*, p. 29). The fact that any actual system of conventional rules will be indeterminate and not prescribe actions in all possible cases is not in my view problematic. I believe that a system of ethics *should be* incomplete in the sense that Hodgson has in mind. Act Utilitarianism is a good example of an ethics which is complete in Hodgson's sense. Since no two actions have identical consequences, Act Utilitarianism dictates a uniquely right action in every situation. Such a moral theory is too confining. Ethical agents should have a significant sphere of action in which many morally permissible options are open to them. Concerning Hodgson's contention that ethical rules merely prescribe the minimal standards of decent behavior, he is mistaken. Actual ethical standards are indeed primarily concerned with the minimum necessary for communal existence, but ethics is not exclusively concerned with these minimums. Actual ethical conventions include "counsels of perfection," which recommend levels of forgiveness, understanding, and generosity far beyond the basic requirements of decency. I think that it is useful for a moral theory to incorporate ethics with respect to both the permissible and the supererogatory. For simplicity's sake, in this paper I focus on moral permissibility. In a more complete moral theory we would also define the supererogatory in a way that would piggyback upon the actual ethical standards which a society holds up for its members as its ethical ideal. I think that by the end of this paper a perceptive reader would have little trouble seeing how the morally praiseworthy could be defined in a manner that parallels what I am doing for the morally permissible.

The difficulty of potential conflicts between conventional rules is more serious and requires a more thorough discussion. Were it not for human fallibility the almost infinite complexity of human life would argue against rule-based morality. The same fallibility which compels human societies to rely on rules necessitates that the rules be simple. The simplicity of the rules relative to the complexity of life creates two problems: there will arise exceptional circumstances in which it would be better not to follow the rule and there will be cases in which multiple rules apply and conflict. The Rule Utilitarian must explain and justify exactly how she deals with these related problems. A sensible middle-of-the-road position between treating rules merely as rules of thumb and Kantian rigidity is desirable. Too much flexibility opens the gates to rationalization. Excessive rigidity comes at such a cost as to undermine respect for the rules. The experience of living together has shown the need for ethical rules that are mostly insensitive to consequentialist considerations but which do allow exceptions in extreme circumstances.

Two considerations jointly yield the correct delineation of the reasonable permissible exceptions to rules. The first step is to include a general disaster avoidance qualification.[19] I am tempted to assert that no modification of ARU2 is needed here because a general disaster avoidance qualification is built into the actual ethical conventions of real human societies. People are not such fools as to fail to understand that on occasion strict adherence to a rule might lead to disaster and as a result they typically understand ethical principles as implicitly including disaster avoidance clauses. But there is no harm in having our moral theory require that disaster avoidance clauses be built into legitimate ethical rules. The second step is to note that in real ethical systems there is often a well-recognized hierarchy of rules such that it is generally understood that some rules take precedence over others. One should not lie. This is surely a common ethical convention. One should also protect innocent people from being murdered. The ethical conventions of our society surely place greater importance on the rule about protecting the innocent. In cases of conflict the second rule takes priority over the first. Appealing to the ranking of rules within the actual ethical conventions of a society is thus an important resource for the moral theorist to exploit. This resource has its limits, however. It is unlikely that any actual ethical systems clearly rank all rules into a well-ordered hierarchy. For example, the rule against lying and the rule against breaking a promise are clearly distinct rules for our ethical community. Yet I think neither of these rules clearly outranks the other. Both concern integrity in our dealings with others and, in the rare instances where the promise collides with our obligation to tell the truth, I think serious and reflective individuals are unsure of which rule has priority.

---

[19] See the works of Hooker previously cited in note 10.

The difficult cases are those in which two or more legitimate rules prescribe contradictory courses of action without any clear priority of one of the rules. Real ethical systems are not so finely calibrated as to anticipate every possible situation, indeed given the cognitive limitations of human beings, could not be. So, realistically formulated rules are bound to conflict at times. In such cases the system of actual rules gives no clear guidance to the conscientious individual. Rule Utilitarians, I have already mentioned Johnson in this context but Mill himself is also an example, often resort to some version of the greatest happiness principle to fill in gaps or to repair inconsistencies in a system of rules. While I respect precedent, I think it is preferable simply to consider all actions not clearly and consistently proscribed by the system of rules to be morally permissible actions. I suggest the following: when two or more legitimate and equally significant rules give conflicting commands, the agent is blameless if he obeys either rule. Whatever is not consistently forbidden is permitted.

It may be helpful to the reader if I explain how these two points resolve questions about some standard cases of possible exceptions to rules. The usual objection to Kant's demand for exceptionless truth-telling is that it is surely unreasonable to think that it is wrong to lie to a murderer in order to protect his proposed victim. This result is easily reached by means of the disaster avoidance qualification. The murder of an innocent person is surely serious enough to count as a disaster. Contrast this with the Desert Island Promise. If I have promised a dying man to perform some minor but unpleasant task, then the fact that the consequences of keeping the promise are less pleasant than the consequences of breaking the promise does not constitute a reasonable exception. Unpleasantness is not disaster, thus I am required to keep my promise. The rule does not allow that much flexibility. Finally, if the only way to keep a promise is to tell a lie, then whichever course of action I choose I am blameless. The reader will note that these standard examples, often used as putative counterexamples to test the validity of moral theories, are here being used merely to illustrate the meaning of the theory by specifying its implications. This illustrates a proper use of thought experiments from the point of view of the pragmatic revisionary method.

At this point the number of explanations and qualifications of the theory are so numerous that to build all of them into the formal statement of the position would make it very hard to read. Accordingly I will introduce three terms that will allow me to reformulate Actual Rule Utilitarianism more concisely. A "legitimate ethical rule" is an ethical practice that is de facto in a given society, contains a disaster avoidance qualification, and promotes the happiness of people affected by the rule. To promote happiness a rule does not have to be ideal. It merely has to make life better than it would be otherwise. One rule has "greater weight" than another rule if both rules are de facto in a given society and if mature individuals in the society generally agree that the first rule takes precedence over the second. Finally, two rules are of "equal weight" if mature individuals in the society do not generally agree that one of the rules takes precedence over the other. These terms allow the final reformulation of Actual Rule Utilitarianism incorporating the necessary qualifications.

(ARU3) An act is morally right either if it is not proscribed by any legitimate ethical rule in the society in which it would occur, or if so proscribed it is also required by another legitimate ethical rule of equal or greater weight.

I hope by this point to have explained why a moral theorist such as myself who is conservative enough to have considerable respect for the ethical conventions of real human societies but who nevertheless wants to take a critical and reflective stance towards them would be attracted to Actual Rule Utilitarianism. Since most moral theorists are more critical of ethical conventions than I am, it is wisdom on my part to attempt to explain that the theory is not excessively conservative. The central idea of Actual Rule Utilitarianism is that responsible people are morally obliged to conform their actions to the legitimate ethical rules of their community and that this is all that they are morally obliged to do. In view of the similarity of my theory to Conventionalism it seems appropriate to review the standard objections to the latter as a way of highlighting the more progressive nature of the former.

One minor objection is this: What does the theory say about two human beings who are complete strangers to each other, who have no shared membership in any ethical community? What ethical obligations do they owe each other? My answer is simple though it may seem too harsh to some delicate minds. Each has whatever obligations their ethical community says they have towards strangers. Should any human beings find themselves in a perfect state of nature, being members of no ethical community whatsoever, they would have no ethical obligations at all. The harshness of this answer is not a fault of the theory; it is a reflection of the harshness of the state of nature. And it is this very harshness that motivates any rational human being to seek to join or create some ethical community.

Another classic objection to Conventionalism is that individuals seem to participate simultaneously in multiple overlapping ethical communities. The same person may be a member of a professional organization, a civic club, a church, a neighborhood, and an ethnic group, as well as being a citizen of her nation-state. Which of these groups constitutes her society for the purposes of identifying her moral duties? Such a question may as legitimately be directed against the position I have articulated here. Actual Rule Utilitarianism can give a simple and direct answer to this challenge: the agent is obligated to conform to all the legitimate moral rules of all the "societies" to which she belongs. The reason that ARU3 has no problem with such an inclusive answer is that it already has built-in a method of resolving conflicts between rules. Thus potential conflict between the rules of different "societies" to which the agent belongs is not a problem. One additional point concerning the multiple affiliations of individuals should be noted. Small subgroups in a larger society often create rules that are not morally legitimate. Professional organizations, many people suspect, sometimes create rules designed to protect their financial interests at the expense of the society at large. Such rules, while perfectly rational in their own way, would fail to promote the general welfare and so fail morally to bind their members.

Conventionalism accepts whatever ethical rules societies have in place and holds the individual morally accountable to those rules. What conventionalism lacks is a way of critically reflecting on those rules and rejecting the legitimacy of those which should be rejected. In general it seems too conservative to most philosophers' taste. Utilitarianism, on the other hand, has always been a progressive doctrine. Bentham and Mill were attracted to it in large part because it seemed to them to have implications for the improvement and enlightenment of mankind. Far from approving of all the institutions of their society, the classical Utilitarians were attracted to the greatest happiness principle precisely because it seemed to them to provide firm ground from which to argue for the reformation of society. Does Actual Rule Utilitarianism betray that progressive spirit? Like conventionalism, which it strongly resembles, it favors the status quo. It holds that we are obliged to obey the actual moral rules of our society even when these are imperfect. The only duties which one needs to consider are those already recognized by society. It departs from Conventionalism in that it uses the principle of utility to test those rules, rejecting the moral authority of those which fail the test. Is this too conservative?

Actual Rule Utilitarianism is modestly progressive not utopian. It is progressive in that it provides a basis to criticize pernicious institutions like slavery or female genital mutilation. It does not, however, insist on perfection or even place a moral obligation on individuals actively to reform their society. (That is unless some society has an actual rule requiring individuals to be reformers, which I doubt.) But there is no rule which forbids a thoughtful person from reflecting on his society and carefully considering which rules and institutions are good and which are bad, and of the good, which can be improved. Nor would such a rule be legitimate if it did exist. Moreover, societies could, and I believe our society actually does, encourage thoughtful citizens to reflect on its institutions and to engage in thoughtful dialogue about how we might improve. Society treats the sincere and well-considered attempt to reform society as supererogatory and not as a general moral requirement.[20] Indeed Mill's career in Parliament championing reform is an example of the pursuit of social progress and reform. I think it best to consider such activism praiseworthy but not a moral requirement falling on people in general.

In conclusion I will consider the issue of slavery in more detail in order to sharpen the contrast between Actual Rule Utilitarianism and conventionalism. Actual Rule Utilitarianism, unlike conventionalism, accepts only those rules that promote human happiness as morally authoritative. The standard Utilitarian view of slavery is that it does not promote human happiness, indeed that it is severely dysfunctional, and hence wrong. This classic Utilitarian answer is displeasing to more rationalistically inclined philosophers who want proof that slavery is necessarily wrong and are dissatisfied with any argument depending on "mere" empirical fact. My response to such objections is no different from that of most Utilitarians: the search for logically necessary moral truths is a chimera. This answer will not satisfy the Kantians but there it is. Slavery is an institution which does not promote human happiness. This is not a necessary truth but it is nevertheless true. Hence rules permitting slavery, since the institution does not promote the general welfare, are not morally binding.

---

[20] Of course there is considerable disagreement about who is really sincere and who is appropriately thoughtful, but that is just to be expected. Such disagreements are quite noticeable while a social reform is being argued over. But in retrospect society praises those who have challenged us to do better.

Unfortunately, there is reason to suspect that this standard utilitarian answer may not work for the Actual Rule Utilitarian. I argued above that right actions are actions which are not forbidden by legitimate rules. The generous way in which I have defined the morally permissible poses a problem here. Actions are right unless there is a legitimate rule proscribing them. So, for keeping slaves to be wrong there must actually be a legitimate ethical rule in the society which forbids treating people as slaves. It does not seem likely that a society which approves of slavery could also have a legitimate ethical rule forbidding it. Hence, it seems that Actual Rule Utilitarianism cannot condemn the keeping of slaves in societies which permit slavery. This result, should it hold up, really does seem too conservative and too much like conventionalism to be acceptable. Can Actual Rule Utilitarianism make a stronger case against slavery than this?

Paradoxical as it sounds, I will argue that societies which have ethical rules permitting slavery also typically have rules forbidding it. I have found much to praise in Johnson's understanding of ethics on analogy with law. However, it is the disanalogies between ethics and law to which I now want to call attention. Law, for the most part, consists of a system of rules which are as fully explicit as society can make them. Only one formulation of a given law is taken to be official and we appoint judges as official interpreters of the law. The professional guardians of our judicial system work hard to minimize vagueness and keep the system consistent. Recent psychological research has verified what most of us always suspected: ordinary people navigate their world with fuzzy concepts. The standards of clarity which prevail in science, philosophy, and the law are the exception in human life, not the norm. I suspect that normal thinking about ethical matters is one of the most fluid and vague areas of human life. The structure of the ethical rules of a community is much less precise and more holistic than that of the law. The informal nature of ethical systems makes it a difficult matter to tease out one part of that system in order to subject it to criticism while accepting the whole of which it is an integral part. There is no official codification of ethical convention to consult as there are statute books of the law. There is a common understanding of what ethical behavior looks like but it is not precise, and it admits of considerable scope for individual variation. Such imprecision would be intolerable in a legal system which necessarily includes mechanisms for formal punishments, but in ethics, where the only punishment consists of the disapproval of our fellows this is not so much of a problem. The imprecision of ethical rules fits well with their decentralized status. In the law we appoint particular individuals to legislate, others to enforce, still others to judge. In an ethical community each individual determines what the rules really are, who has violated them and how much disapproval is warranted when they do. We each do this individually and in dialogue with others. Informal ethical deliberation makes up a significant part of people's casual conversations. By what is usually a much slower process individual members of the ethical community participate in the process of revising our shared understanding of what it means to be a good person.

Since the system is so informal it is impossible to delineate precisely where one rule ends and another begins. Within a typical ethical system there are rules that concern how people in general are to be treated and a series of subrules or qualifications which modify the general rule. For example, we think that people ought to be free to come and go as they please. We think they have a right to speak their mind. We think they should be paid for their work. These are all ethical principles of wide generality. But we also think that children are exceptions to these rules. It is quite proper to restrain children from coming and going as they please, from speaking their mind at will, and there is no need to pay them for routine household chores. In slave societies slaves were also considered different enough from "normal" people to be treated differently. As part of the ideology of slavery, it was believed that slaves, like children, could not adequately care for themselves so "normal" people could properly impose their will on slaves. Slavery gradually ceased to be an accepted institution within the ethical norms of society when thoughtful people reflected upon the nature of the ethical system in place in their society and came to the realization that the distinction between slaves and free persons was a serious and fundamental mistake. How did people in fact come to see that slavery was wrong? The story of John Woolman colorfully illustrates the slow, informal process. Woolman, being one of the few literate people in colonial America, was helping his neighbor compose a will. Upon listing the man's property Woolman was suddenly struck by the fact that along with the rest of the material possessions of his neighbor he was being asked to record the names of people that he knew. People who happened to be slaves. Woolman was so struck by the recognition that he put down his pen. This eventually lead Woolman to take up the cause and to travel about colonial America talking to individual slave owners trying to convince them to stop treating these people as a separate inferior class of human beings. Woolman's efforts did lead in relatively short order to a complete rejection of slavery by his fellow Quakers, but it took several more generations for the abolitionist movement to convince the nation as a whole to abolish the distinction between slave and free. Mill's language in the final chapter of *Utilitarianism* echoes the idea that large scale progress primarily consists in abolishing invidious ethical distinctions. "The entire history of social improvement has been a series of transitions, by which one custom or institution after another, from being supposed primary necessity of social existence, has passed into the rank of a universally stigmatized injustice and tyranny. So it has been with the distinctions of slaves and freemen, nobles and serfs, patricians and plebeians; and so it will, and in part already is, with the aristocracies of colour, race, and sex."[21] From the point of view of ARU3 this is exactly the sort of ethical reflection that societies should constantly be undergoing, purging the worse parts from our ethical systems. It is the process by which ethical rules come to be rejected, revised, or replaced. It is not a tidy process that appeals to philosophical purists, but it is a realistic one that appeals to those of us with a more pragmatic bent.

Ordinary people understand the ethical standards of their community and how they work even though articulating the complex social processes involved is difficult. My revisionary view of philosophical method has not led me to very radical conclusions. What I am suggesting is really a very modest revision to the way ordinary people understand ethics. I propose that we evaluate the ethical rules we try to live by not according to tradition, taste, or conformity with Scripture, but only according to the degree to which the rule promotes our living together in harmony with our fellows. This, I take it, is very much what Mill had in mind.

I began this paper with a call to incorporate ethics within moral theory. Socrates described himself as a gadfly to the Athenians, stinging them to make them think harder and more critically about their conception of the Good. But Socrates did not try to create a vision of the Good wholly separate from that of his fellow Athenians. Instead he always began from where his interlocutors stood and tried to pull them forward one step at a time. Perhaps Socrates was being ironic when he said he hoped to learn some truth by listening to other people but I am not being ironic when I say that philosophers have much to learn about right and wrong by taking ethical traditions seriously. The moral theories we construct should incorporate those traditions as a foundation. Our proper role is to offer improvements to our traditions. This concept of the proper relation between ethics and moral theory has led me to accept a version of Utilitarianism which has not been taken seriously for nearly fifty years. Among the advantages of this theory are: (1) it is simple, (2) the rules are publicly understood, (3) it respects individual freedom, (4) it promotes trust and cooperation, and (5) it allows for progress. I see no other normative moral theory with the same advantages, but I would welcome discussion of other theories along the same lines as those followed in this paper.

---

[21] Chapter V, paragraph 37.

What we should care about is whether or not what philosophers say about right and wrong will help or hinder us in our goal of living together harmoniously and cooperatively. I believe that my understanding of the nature and purpose of moral theory is more consistent with the great and ancient tradition of philosophical thought from Socrates to Mill. Since Mill's time the audience for philosophy has gradually narrowed and philosophy has become mostly a private conversation among the cognoscenti. I retain the small and probably naïve hope that a more pragmatic approach to philosophical method might win some of that audience back.

RICHARD B. MILLER

East Carolina University

WHAT IS LYING?*

It's not a lie if you believe it.
—George Costanza

A lie with a nod is still a lie, but it's an easy lie.
—Philip Marlowe

The main question that we typically have about lying is whether someone (a politician, a witness, a spouse, and so on) is lying to us. And we frequently expend a lot of energy to find out. For example, the police use polygraphs that monitor heart rate, perspiration, and other physiological indicators of lying on criminal suspects. And, recently, researchers have even used functional Magnetic Resonance Imaging (fMRI) to find that different areas of the brain are activated depending on whether someone is lying or telling the truth.[1]

But there are also many important philosophical questions about lying. For example, many moral philosophers have studied the morality of lying.[2] Most notably, Immanuel Kant (*op. cit.*) argued that it is always wrong to lie by asking us to imagine what would happen if everybody lied when it was to their advantage. More recently, Sissela

---

*I would like to thank Jonathan Adler, Jody Azzouni, Matt Bedke, Tom Carson, Jane Carter, Jacob Caton, David Christensen, Tony Doyle, Martin Frické, Jerry Gaus, Allen Habib, Terry Horgan, Peter Lewis, Pen Maddy, James Mahon, Kay Mathiesen, Marc Moffett, Shaun Nichols, Patrick Rysiew, Roy Sorensen, Amie Thomasson, Susan Vineberg, Dennis Whitcomb, and Dan Zelinski for much helpful feedback. I would also like to thank audiences at the Philosophy Department at the University of Arizona, at the Philosophy Department at Wayne State University, at the Department of Logic and Philosophy of Science at the University of California, Irvine, at the Philosophy Club at Rutgers University, at the Philosophy Club at the John Jay College of Criminal Justice, and at the Pacific Division Meeting of the American Philosophical Association.

[1] See, for example, Feroze B. Mohamed, Scott H. Faro, Nathan J. Gordon, Steven M. Platek, Harris Ahmad, and J. M. Williams, "Brain Mapping of Deception and Truth Telling about an Ecologically Valid Situation: Functional MR Imaging and Polygraph Investigation—Initial Experience," *Radiology*, CCXXXVIII (2006): 679–88. There are also techniques that can be used to identify lies in recorded information. For example, researchers have used textual analysis to find that liars are somewhat less likely to use first-person pronouns. See, for example, Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards, "Lying Words: Predicting Deception from Linguistic Styles," *Personality and Social Psychology Bulletin*, XXIX (2003): 665–75.

[2] For example, Plato, *Republic*, C.D.C. Reeve, trans. (Indianapolis: Hackett, 2004 [360 BCE]); Augustine, "Lying," in *Treatises on Various Subjects*, Volume XVI, Roy J. Deferrari, ed. (New York: Fathers of the Church, 1952 [395]), pp. 53–120; Thomas Aquinas, "Of Lying," in *Summa Theologica*, Volume XII (London: Burns, Oates, and

0022-362X/09/0601/29–56

Bok (*op. cit.*) argued that it is wrong to lie more often than we think because we often underestimate the personal and social costs of lying.

In addition, many epistemologists are interested in lying.[3] A great number of our beliefs about the world are based on what other people tell us. But it is not clear whether these beliefs count as knowledge if these other people might be lying to us.

In order to answer such questions about lying in ethics and epistemology, we first need to know what it means for something to be a lie.[4] In other words, philosophers need to define their terms. Saying what it means for something to be a lie is the goal of this paper.

I. CONCEPTUAL ANALYSIS

In order to provide such a definition, I will be engaging in *conceptual analysis*.[5] Plato famously used this technique in his dialogues to try to understand such concepts as justice, knowledge, and love, and it has been an important part of philosophical practice ever since. (Harry Frankfurt's attempt to understand the concept of *bullshit*—by contrasting it with lying—is a notable, contemporary example of conceptual analysis.[6]) However, over the last several years, the legitimacy

---

Washbourne, 1922), pp. 85–98; Immanuel Kant, *Foundations of the Metaphysics of Morals*, Lewis W. Beck, trans. (New York: Macmillan, 1959 [1785]); Sissela Bok, *Lying* (New York: Random House, 1978); Jonathan E. Adler, "Lying, Deceiving, or Falsely Implicating," this JOURNAL, XCIV, 9 (September 1997): 435–52. In fact, this is arguably a central issue in ethics. For example, according to Kant, lying is "the greatest violation of a human being's duty to himself regarded merely as a moral being" (quoted in Bernard Williams, *Truth and Truthfulness* (Princeton: University Press, 2002), p. 102).

[3] For example, Ludwig Wittgenstein, "Notes for Lectures on 'Private Experience' and 'Sense Data'," *Philosophical Review*, LXXVII (1968 [1935]): 275–320, here p. 280; Tyler Burge, "Content Preservation," *Philosophical Review*, CII (1993): 457–88, here p. 474; Peter J. Graham, "What Is Testimony?" *Philosophical Quarterly*, XLVII (1997): 227–32; Jennifer Lackey, "Norms of Assertion," *Noûs*, XLI (2007): 594–626, here p. 602; Paul Faulkner, "On Dreaming and Being Lied to," *Episteme*, II (2006): 149–59; Dan O'Brien, "Testimony and Lies," *Philosophical Quarterly*, LVII (2007): 225–38. In fact, just like Kant, Faulkner recently asked us to imagine what would happen if everybody lied when it was to their advantage. Basically, whereas Descartes was essentially worried about possibly being a character in the film *The Matrix*, Faulkner is worried about possibly being a character in the film *The Truman Show*.

[4] Cf. Thomas L. Carson, "The Definition of Lying," *Noûs*, XL (2006): 284–306, here p. 284.

[5] See, for example, Frank Jackson, *From Metaphysics to Ethics* (New York: Oxford, 1998); Eric Margolis and Stephen Laurence, "Concepts," *Stanford Encyclopedia of Philosophy* (2006): http://plato.stanford.edu/entries/concepts/.

[6] Frankfurt, *On Bullshit* (Princeton: University Press, 2005). Given their close connection, it may be possible to give a definition of bullshit along lines very similar to the definition of lying that I propose in section II below (that is, you are *bullshitting* if you say something for which you (believe you) lack adequate evidence and you believe

of conceptual analysis has been called into question.[7] Thus, it will be useful to clarify exactly what the goals and methodology of this paper will be.

For each concept that he studied, Plato tried to develop a *definition* that correctly classified things as falling under that particular concept or not. Ideally, he tried to identify *necessary* and *sufficient* conditions for whether things fall under that concept (cf. Margolis and Laurence, *op. cit.*, section 2.1). In our case, we would like a definition that rules in everything that is a lie and that rules out everything that is not a lie.[8]

In order to determine whether a proposed definition is correct, however, we need to have fairly reliable intuitions about whether particular (often hypothetical) cases fall under the given concept. Epistemologists, for example, typically test proposed definitions of knowledge in just this way.[9] In our case, we have to be able to tell (given that we know certain things about the beliefs and intentions of the speaker) whether specific statements are lies.

Conceptual analysis has been criticized because of this reliance on intuition. For example, several people have argued that our intuitions about hypothetical cases cannot give us "a priori knowledge of necessary truths" (Melnyk, *op. cit.*, p. 267). Also, there is empirical evidence that such intuitions are fallible.[10]

But even if our intuitions can only provide us with fallible, a posteriori knowledge, there are good reasons to think that they can be used effectively to test proposed definitions of important concepts

---

that you are in a situation where the following norm of conversation is in effect: "Do not say that for which you lack adequate evidence."). However, I will not pursue that possibility further in this paper.

[7] See, for example, Stephen Stich and Jonathan M. Weinberg, "Jackson's Empirical Assumptions," *Philosophy and Phenomenological Research*, LXII (2001): 637–43; Weinberg, Shaun Nichols, and Stich, "Normativity and Epistemic Intuitions," *Philosophical Topics*, XXIX (2001): 429–60; Margolis and Laurence, *op. cit.*, section 5.2, Andrew Melnyk, "Conceptual and Linguistic Analysis: A Two-Step Program," *Noûs*, XLII (2008): 267–91.

[8] The question of what 'lying' means is, of course, different from the question of whether someone is lying to us on a particular occasion. Even if we know what the definition of lying is, we may not know whether the definition is satisfied on any particular occasion (cf. Williams, *op. cit.*, p. 97).

[9] See, for example, Edmund Gettier, "Is Justified True Belief Knowledge?" *Analysis*, XXIII (1963): 121–23.

[10] For instance, our intuitions about hypothetical cases can be influenced by the order in which the cases are presented to us. See, for example, Stacey Swain, Joshua Alexander, and Weinberg, "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp," *Philosophy and Phenomenological Research*, LXXVI (2008): 138–55. Also, our intuitions can be influenced by our moral beliefs. See, for example, Joshua Knobe, "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *Philosophical Studies*, CXXX (2006): 203–31.

(cf. Melnyk, *op. cit.*, p. 288). For instance, our intuitions have to be a pretty good guide to how words are commonly used (cf. Jackson, *op. cit.*). If they were not, we would have a lot of trouble communicating with each other. Furthermore, trying to capture common usage is arguably a good way to identify useful concepts.[11] It would be very surprising if humans had developed terms like 'knowledge' and 'lying', but these terms were not getting at important phenomena in the real world.

Admittedly, people from different cultures may use the same words very differently (cf. Weinberg, *op. cit.*). But the goal in this paper is to determine how the word 'lying' is used in *this* culture. Thus, relying on the intuitions of members of this culture (in this case, philosophers) seems perfectly appropriate.[12]

Even if we have reliable, shared intuitions, however, we might still worry that the goal of conceptual analysis is unattainable. That is, there may not be necessary and sufficient conditions for whether things fall under a given concept. Instead there may simply be prototypical instances of the concept with different things falling closer to or further from these prototypes.[13] However, capturing common usage is only one desideratum for a definition of lying. Since we are looking for a definition for philosophical use, we are also concerned with the simplicity and theoretical utility of the definition. Also, while it may not be possible to capture common usage perfectly with a concise definition of lying, some proposed definitions will do better than others. In this paper, I will offer a definition of lying that captures common usage better than the other proposed definitions.[14] I will argue that the other proposed definitions are either too broad (that is, they rule in some things that are clearly not lies) or too narrow (that is, they rule out some things that clearly are lies).

---

[11] Cf. J.L. Austin, "A Plea for Excuses," *Proceedings of the Aristotelian Society*, LVII (1956): 1–30, here p. 8. This is not to say that conceptual analysis is the only way to identify useful concepts. Nor is it to say that conceptual analysis is guaranteed to identify useful concepts.

[12] Even within the same culture, different people may have conflicting intuitions about particular cases (cf. Carson, *op. cit.*, p. 301). This might indicate that there is more than one concept in play. However, while there are certainly some conflicting intuitions about lying (cf. footnote 17 below), there seems to be fairly broad agreement on the central cases. Even so, it would be useful to perform an empirical study (à la Weinberg et al., *op. cit.*, or Swain et al., *op. cit.*) to test the proposed definitions against intuitions of people who do not have a stake in this philosophical debate.

[13] See, for example, Stich and Weinberg, *op. cit.*, pp. 638–40; Marjorie Taylor, Gretchen L. Lussier, and Bayta L. Maring, "The Distinction between Lying and Pretending," *Journal of Cognition and Development*, IV (2003): 299–323; Margolis and Laurence, *op. cit.*, section 2.2.

[14] It turns out that most philosophers who discuss lying are actually concerned with something more restrictive than 'lying' as the term is commonly used. In the final section of this paper, I will argue that a slight modification to my definition of lying can capture this narrower concept.

II. MY DEFINITION OF LYING

There is an informal tradition in the conceptual analysis literature of sneaking up on the official definition of some concept. That is, one starts with an obvious candidate definition, says why it will not quite work, tweaks the definition, says why it still will not work, tweaks the definition again, and so on. By the time the reader reaches the official definition, she is often dizzy enough to capitulate. I am going to break with tradition and simply start out by stating my official definition.

I think that you lie when you *assert* something that you believe to be false.[15] Basically, you lie when you "go on the record" with something that you believe to be false (cf. Carson, *op. cit.*, p. 290, Sorensen, *op. cit.*, p. 252).

You *lie* to *X* if and only if: (*AL*)

(1) You assert that *p* to *X*.
(2) You believe that *p* is false.

Of course, in order to really define lying, I should also say exactly what it means to assert something. I think that you *assert* something when (a) you say something and (b) you believe that you are in a situation where you should not say things that you believe to be false. More precisely, you assert something when you say something and you believe that Paul Grice's first maxim of *quality* (namely, "Do not say what you believe to be false") is in effect as a *norm of conversation*.[16]

Plugging this account of assertion into *AL*, we get my definition of lying.[17]

---

[15] Cf. Roderick M. Chisholm and Thomas D. Feehan, "The Intent to Deceive," this JOURNAL, LXXIV, 3 (March 1977): 143–59, here p. 152; Adler, *op. cit.*, p. 435; Gary Watson, "Asserting and Promising," *Philosophical Studies*, CXVII (2004): 57–77, here p. 72; Roy Sorensen, "Bald-Faced Lies! Lying without the Intent to Deceive," *Pacific Philosophical Quarterly*, LXXXVIII (2007): 251–64, here p. 256.

[16] Grice, *Studies in the Way of Words* (Cambridge: Harvard, 1989), p. 27. You can deceive people by violating several different norms of conversation (cf. Grice, p. 30; D.S. Mannison, "Lying and Lies," *Australasian Journal of Philosophy*, XLVII (1969): 132–44, here p. 132). But you need to violate Grice's first maxim of quality, in particular, in order to be lying. I am assuming that, if this norm is in effect and you do not act in accordance with it, then you have violated it. Thus, a simpler statement of my definition would be that you are lying if you believe that you are violating Grice's first maxim of quality.

[17] In this paper, I argue that my definition of lying gets all of the central cases right. But there are many borderline cases where intuitions are less certain (cf. Carson, *op. cit.*, p. 294; Sorensen, *op. cit.*, pp. 258–59). In several footnotes below, I describe how my definition can easily be modified—in a way that preserves the essential insight that lying has to do with violating a particular norm of conversation—to accommodate

You *lie* to *X* if and only if: (*BNL*)

(1) You state that *p* to *X*.
(2) You believe that you make this statement in a context where the following norm of conversation is in effect:

Do not make statements that you believe to be false.

(3) You believe that *p* is false.

This definition of lying correctly rules in many prototypical instances of lying. For example, suppose that I am at a fancy party in Washington, D.C., and a beautiful woman comes up and asks who I am. I might say with complete seriousness, "I am the Prince of Denmark" in order to impress her.[18] My definition gives the intuitively correct result that this is a lie. In this case, I make a statement, I believe that it is false, and I believe that this is a situation where I should not say what I believe to be false. Similarly, Silvio is lying if he testifies in court that "Tony was home with me at the time of the murder" in order to deceive the jury and get his boss acquitted.[19]

In addition, my definition correctly rules out many false statements that are not lies. For example, suppose that I am on stage at my local community theater and a beautiful woman comes up and asks who I am. I might then say that "I am the Prince of Denmark" because it is a line from the play that we are performing. My definition gives the intuitively correct result that this is not a lie. In this case, I make a statement, I believe that it is false, but I do not believe that this is a situation where I should not say what I believe to be false.[20] Similarly,

---

conflicting intuitions on such cases. For example, if Graham Priest states that "This sentence is false" in a situation where he believes that Grice's first maxim of quality is in effect, it is not clear that he is lying. Priest believes that this statement is false *and* that it is true. See his "The Logic of Paradox," *Journal of Philosophical Logic*, VIII (1979): 219–41. Thus, he is obeying the maxim "Try to make your contribution one that is true" (Grice, *op. cit.*, p. 27). In order to accommodate the intuition that Priest is not lying, my definition might be modified to use the norm: *Do not make statements that you believe to be only false.*

[18] I looked him up on Wikipedia. He is about the same age and he looks a little bit like me. So, I might get away with it. But I am lying regardless of whether I expect to get away with it.

[19] Grice's maxims are usually unstated. But when Silvio takes the witness stand and swears to tell the truth, the whole truth, and nothing but the truth, he makes it fairly explicit that he believes that Grice's first maxim of quality is in effect. When he then lies under oath, it counts as *perjury* (cf. Stuart P. Green, *Lying, Cheating, and Stealing* (New York: Oxford, 2006), pp. 133–47).

[20] It might be suggested that an actor on stage is not making *statements* at all. But for purposes of this paper, I do not draw any distinction between *stating* something

Silvio is not lying if he says to Paulie at the club, "Tony was home with me at the time of the murder" and then winks to indicate that he is not to be taken seriously. By winking, Silvio *turns off* (or "opts out from the operation" of) Grice's first maxim of quality with respect to this particular statement (cf. Grice, *op. cit.*, p. 30, Watson, *op. cit.*, p. 65).[21]

III. NORMS OF CONVERSATION

Before I discuss why the other definitions of lying are wrong, there are a few important things that should be noted about the assertion condition in my definition. First, my analysis of assertion in terms of Grice's first maxim of quality is not intended to be a *complete* analysis of assertion. I am only trying to capture the normative component of assertion that is necessary for lying.[22]

Second, the assertion condition is not that you believe that you are in a situation where, *all things considered*, you should not say things that you believe to be false. For example, suppose that a homicidal maniac shows up in the audience of an improvisational performance piece, demands that the performance continue, but threatens to shoot any performer who says something that she believes to be false. This is certainly a situation where the performers believe that they should not say things that they believe to be false. Since this is a performance, however, the performers do not believe that Grice's first maxim of quality is in effect as a norm of conversation. Thus, the performers would not be lying if they did disobey the maniac.

---

and just *saying* something (cf. Chisholm and Feehan, *op. cit.*, p. 151). It should be noted that actors *can* lie on stage (that is, when they speak *sotto voce* to each other). But utterances that are part of the performance are never lies.

[21] Even if a speaker sends such a signal, the audience may not pick up on it (cf. Carson, *op. cit.*, p. 294). For example, an adult might tell a tall tale (that is, an exaggerated story) to a group of very young children that he expects to take him totally seriously. Although it is clear to the speaker that Grice's first maxim of quality is not in effect in this situation (in fact, telling children such tales ultimately helps them to learn when this norm is not in effect), some (for example, Carson, *op. cit.*, p. 298) might want to say that he is lying. In order to accommodate that intuition, my definition might be modified to simply require that a liar believe that *his audience believes* that the norm is in effect. Also, if a visitor to a foreign country does not know for sure whether a wink turns off Grice's first maxim of quality in this unfamiliar cultural context, some might want to say that she is still lying even if she winks after saying something that she believes to be false (especially if she intends to deceive). In order to accommodate that intuition, my definition might be modified to simply require that a liar *fail to believe* that the norm is *not* in effect.

[22] More substantive accounts of assertion have been proposed (cf. Robert Brandom, "Asserting," *Noûs*, XVII (1983): 637–50; Watson, *op. cit.*). In fact, some philosophers (such as Timothy Williamson, "Knowing and Asserting," *Philosophical Review*, CV (1996): 489–523) have even suggested that the relevant norm is: *Do not make statements that you do not know to be true.*

Third, the assertion condition is not simply that you believe you are in a situation where a norm is in effect that you should not say things that you believe to be false. For example, suppose that a powerful and beloved monarch does not like to hear false statements even if they are said in jest. As a result, it is considered impolite to say things that you believe to be false in her presence. So, if I say to her that "I am the Prince of Denmark" with a wink, I am violating a norm with the same content as Grice's first maxim of quality. But I am not lying. Because of the wink, Grice's first maxim of quality is not in effect *as a norm of conversation* (and I do not believe that it is in effect) with respect to this particular statement.[23]

The notion of a *norm of conversation* is essentially taken directly from Paul Grice (*op. cit.*, pp. 22–40). In addition to not saying what you believe to be false, you should "not say that for which you lack adequate evidence," you should "make your contribution as informative as is required," you should "avoid ambiguity," and so on. The function of these norms (unlike norms of politeness, for example) is to facilitate communication in normal conversations.[24] As Grice puts it, "all sorts of other maxims (aesthetic, social, or moral in character), ... are also normally observed by participants in talk exchanges ... The conversational maxims, however, ... are specially connected ... with the particular purposes that talk ... is adapted to serve and is primarily employed to serve" (*op. cit.*, p. 28). In such contexts, we (typically) obey such norms, expect other people to obey them, and believe that other people expect us to obey them. We quickly learn when such norms are in effect (for example, in a normal conversation) and when they are not (for example, in the performance of a play). And we have such knowledge even if we have not read Grice and could not put our knowledge explicitly into words.[25] Also, when we believe that such a norm is in effect, we think that it is reasonable to rebuke someone for violating it.[26] But we recognize that such violations are sometimes excusable (namely, when there are other interests or norms that trump the norms of conversation).[27]

---

[23] The norm would typically be back in effect for any subsequent statements.

[24] Social norms typically provide a solution to some interaction problem (cf. Edna Ullmann-Margalit, *The Emergence of Norms* (New York: Oxford, 1977); Philip Pettit, "*Virtus Normativa*: Rational Choice Perspectives," *Ethics*, C (1990): 725–55). For example, the function of norms like "Drive on the right side of the road" is to keep us from running into each other. Norms of conversation, in particular, help us with the coordination problem that we face when we try to communicate with each other. And these norms apply in the case of written as well as oral communication.

[25] Thus, it is clearly possible for young children, as well as adults, to lie on my definition.

[26] In this regard, my account of assertion is very similar to accounts of assertion in terms of *guaranteeing* or *warranting* (see section VI.2 below).

Finally, I am not taking any position on the *moral* force of Grice's first maxim of quality. It could be that you have a *prima facie* moral duty to follow this norm of conversation (as Kant would probably say) or it could have no more force than a rule of etiquette. In fact, it could have no moral force at all. This agnosticism has the advantage of leaving open the question of whether lying is always wrong (cf. Carson, *op. cit.*, p. 288).

IV. POINTS OF GENERAL AGREEMENT

Despite its initial plausibility, no one (as far as I know) has proposed this particular definition of lying (*BNL*). In fact, the other proposed definitions are quite different in many respects. And, as we go along, we will consider several examples that my definition gets right and that the other definitions do not. But before we get to the points of disagreement with these other definitions (and why they are wrong), it will be useful to note a few things about the points of agreement.

The first thing that everyone essentially agrees on is that, in order to lie, you have to make a statement.[28] Basically, you have to use language to express a proposition.[29] So, for example, while a poker player who makes a very large bet may be trying to deceive you about the strength of his hand, he is not lying to you. Similarly, someone who packs his luggage to mislead you into believing that he is leaving on a trip has not lied to you.[30]

---

[27] While the norms of conversation are clearly distinct from other types of norms, they are similar in many respects to the social norms that govern other everyday activities. For example, we usually know when the norm "Drive on the right side of the road" is in effect (for example, on a city street) and when it is not (for example, at a demolition derby). We typically rebuke people who violate the norm; but we recognize that such violations are sometimes excusable (for example, when a person or a big rock is lying in the road). See Pettit (*op. cit.*) for a discussion of the defining characteristics of social norms.

[28] It is important to note that the statement that you make is not always the same as the literal meaning of your words. For example, suppose that you have a Ph.D. in Philosophy, but have no medical training. If you come across a seriously injured person on the street and say "I cannot help you; I am not a doctor," you are not lying. And if you say "I can help you; I am a doctor," you *are* lying. (Thanks to Jonathan Adler for this example.) The same issue arises with pretending (such as when my niece turns on a flashlight and threatens me with her "light saber") and with metaphor (such as, when I say of my messy roommate that he is a "pig"). Basically, the liar has to believe that her audience understands her statement to mean a certain thing, which she believes to be false. See Chisholm and Feehan (*op. cit.*, pp. 150–51) for an analysis of exactly what it means to make a statement. But since this is a difficult issue that all of the proposed definitions of lying must address, I will set it aside for the purposes of this paper.

[29] You may not have to succeed in actually communicating this proposition to your audience (cf. Carson, *op. cit.*, p. 299). For example, you can probably lie to someone who fails to understand what you are saying or who already believes what you have to tell him.

Admittedly, the word 'lie' is sometimes used to refer to cases where no statement is made. For example, there are "lies of omission" where you fail to state that *p* when it would be expected that you would have if *p* were true. But most philosophers agree that such deceptions are not lies, strictly speaking.[31]

It should be noted, however, that you do not have to make this statement out loud.[32] For example, you can lie by writing something down (such as in a letter). You can lie by sending smoke signals. In fact, as Philip Marlowe points out, you can even lie just by nodding in response to a question.

The second thing that everybody pretty much agrees on is that, in order to lie, you have to believe that your statement is false. So, as George Costanza points out, if you make a statement that you believe to be true, you are not lying.[33] In fact, you are not lying even if you intend to deceive someone by making this statement.[34]

In addition, almost all philosophers think that your statement does not actually have to be false.[35] As Saint Augustine put it, "a person is to be judged as lying or not lying according to the intention of his own mind, not according to the truth or falsity of the matter itself" (*op. cit.*,

---

[30] Cf. Brian Huss, "Bluffing, Lying, and Bullshitting," in Eric Bronson, ed., *Poker and Philosophy* (Chicago: Open Court, 2006), pp. 127–37, here pp. 127–30; Kant, "Ethical Duties towards Others: Truthfulness," in *Lectures on Ethics*, Louis Infield, trans. (New York: Century, 1930), pp. 224–35, here p. 226.

[31] Cf. James Mahon, "The Definition of Lying and Deception," *Stanford Encyclopedia of Philosophy* (2008): http://plato.stanford.edu/entries/lying-definition/, section 1.1. But contrast Paul Ekman, "Lying and Deception," in Nancy L. Stein, Peter A. Ornstein, Barbara Tversky, and Charles Brainerd, eds., *Memory for Everyday and Emotional Events* (Mahwah, NJ: Lawrence Erlbaum, 1997), pp. 333–47, here pp. 334–35. Such "lies" deceive by violating one of Grice's (*op. cit.*, p. 26) maxims of *quantity*. In other words, the speaker is not being "as informative as is required." Although it is not strictly speaking a lie, if you have an obligation to state that *p* to *X* if *p* is true, a lie of omission can be a form of deception that is as morally and epistemically problematic as a lie.

[32] Cf. Frederick A. Siegler, "Lying," *American Philosophical Quarterly*, III (1966): 128–36, here p. 128; Chisholm and Feehan, *op. cit.*, p. 149; Carson, *op. cit.*, p. 287; Mahon, *op. cit.*, section 1.1.

[33] If you innocently repeat a false statement made by someone else, we may call the statement itself a lie. But you are not lying when you pass it along (cf. Harry Frankfurt, "Reply to G. A. Cohen," in Sarah Buss and Lee Overton, eds., *Contours of Agency* (Cambridge: MIT, 2002), pp. 340–44, here p. 340).

[34] Cf. Augustine, *op. cit.*, pp. 56–59; Adler, *op. cit.*; Mahon, *op. cit.*, section 1.2. But contrast Barry O'Neill, "A Formal System for Understanding Lies and Deceit" (2003): http://www.sscnet.ucla.edu/polisci/faculty/boneill/bibjer5.pdf. Although it is not strictly speaking a lie, such a *false implicature* can be a form of deception that is as morally and epistemically problematic as a lie (cf. Adler, *op. cit.*).

[35] See, for example, Siegler, *op. cit.*, pp. 130–32; Mannison, *op. cit.*, pp. 134–37; Warren Shibles, "A Revision of the Definition of Lying as an Untruth Told with Intent

p. 55). In other words, what is relevant is the state of mind of the liar rather than the actual state of the world. For example, in a short story by Jean-Paul Sartre, Pablo Ibbieta tried to deceive the authorities by telling them that Ramon Gris "is hidden in the cemetery."[36] Pablo believed otherwise, but it turned out that Ramon was indeed hiding in the cemetery. Even though his statement was true, Pablo was still lying.

V. INTENDING TO DECEIVE

It is clear that making a statement that you believe to be false is not sufficient for lying. For example, you are not lying if you follow such a statement with a wink or if it is part of a play that you are performing. But there is disagreement among philosophers about how to rule out such cases. In this section, I will discuss the standard strategy for dealing with these cases. In the next section, I will discuss a family of alternative strategies. I argue that these strategies lead to definitions of lying that (unlike my definition) get several cases wrong.

In order to rule out such cases, most philosophers require that you also intend to deceive someone with this statement.[37] In other words, you are lying when you make a statement that you believe to be false with the intent to deceive.

You *lie* if and only if: (*IDL*)

(1) You state that *p*.
(2) You believe that *p* is false.
(3) You intend to deceive by making this statement.

---

to Deceive," *Argumentation*, II (1988): 99–115, here p. 101, Mahon, *op. cit.*, section 1.2. Kant also seems to think that lies do not actually have to be false. See his "On a Supposed Right to Lie from Altruistic Motives." in *Critique of Practical Reason and Other Writings in Moral Philosophy*, Lewis W. Beck, trans. (Chicago: University Press, 1949 [1797]), pp. 346–50. Part of the reason that he thinks that it is wrong to lie to a murderer at the door is that your statement might turn out to be true. As far as I know, Carson (*op. cit.*, p. 285) is the only philosopher who thinks that actual falsity is required for lying. (But he does note that this requirement can easily be dropped from his definition.) Dictionary definitions of lying often do include this requirement. But there can easily be a difference between the way that a word is commonly defined and the way that it is commonly used. In any event, it would be a simple matter to modify any proposed definition of lying, including my own, to accommodate the intuition that actual falsity is required (cf. Carson, *op. cit.*, p. 285).

[36] Sartre, *The Wall* (New York: New Directions, 1948).

[37] See, for example, Augustine, *op. cit.*, p. 56; Mannison, *op. cit.*, p. 133; Bok, *op. cit.*, p. 13; Williams, *op. cit.*, p. 96; Mahon, *op. cit.*, section 1.4. (It should be noted that Augustine (*op. cit.*, p. 60) was not completely sure whether this condition was necessary for lying.) Social scientists (such as J.A. Barnes, *A Pack of Lies* (New York: Cambridge, 1994), p. 11; Ekman, *op. cit.*, p. 334) also typically include this requirement in their definitions of lying. It is also part of most dictionary definitions of lying (cf. Carson, *op. cit.*, p. 286). According to this view, you do not have to actually *succeed* in deceiving someone in order to be lying. You just have to intend to do so.

As it stands, however, this standard philosophical definition of lying is too broad. That is, it rules in some things that are not lies. For example, suppose that I decide to try to convince my new acquaintance at the fancy party that I am an actor (rather than royalty). So, I take a theatrical pose and intone, "I am the Prince of Denmark." Although I have made a statement that I believe to be false with the intent to deceive her (about my occupation), I have not lied to her (for example, about being royalty). I have only *falsely implicated* that I am an actor.[38] Thus, as Bernard Williams (*op. cit.*, p. 96) points out, the definition needs to require that I intend to deceive with respect to the very statement that I make.

In addition, suppose that I make a false statement to a companion (who is in on my little scheme) with the goal of deceiving someone that I know is eavesdropping on our conversation.[39] Since I am not even speaking to this third person, it is not clear that I can be lying to this person. Thus, the definition needs to require that I make the statement *to* the person that I intend to deceive.[40]

You *lie* to X if and only if: (*IDL**)

(1) You state that p to X.
(2) You believe that p is false.
(3) You intend to deceive X with respect to p by making this statement.

---

[38] Cf. Grice, *op. cit.*, pp. 30–31; Adler, *op. cit.* Mark Twain refers to this sort of deception as a "modified lie." See his "My First Lie, and How I Got Out of It," in *The Man That Corrupted Hadleyburg* (New York: Oxford, 1996 [1899]), pp. 167–80, here pp. 173–74. Such a "lie" may be as morally objectionable as lying (cf. Twain, *op. cit.*; Adler, *op. cit.*). It is not lying, however.

[39] Cf. Chisholm and Feehan, *op. cit.*, p. 156. This sort of thing certainly happens in Shakespearean plays and soap operas. For example, in Act II, Scene III of "Much Ado about Nothing," several men hold a loud conversation near where they know Benedick is hiding in order to mislead him into thinking that Beatrice loves him. But it also occurs in real life. For example, fake radio transmissions were sent by the Allies to deceive the Germans during World War II. See J.B. Bell, "Toward a Theory of Deception," *International Journal of Intelligence and CounterIntelligence*, XVI (2003): 244–79, here p. 249.

[40] In a similar vein, my definition of lying rules out the eavesdropper case by requiring that you assert p *to* X. Elizabeth Fricker claims that assertions cannot be targeted at particular individuals. See her "Second-Hand Knowledge," *Philosophy and Phenomenological Research*, LXXIII (2006): 592–618, here p. 598. That is, anyone within earshot can acquire knowledge from a sincere assertion. This might seem to suggest that I am lying to the eavesdropper. But assertions do have to have an intended audience (cf. Fricker, *op. cit.*, p. 596). And, in the eavesdropper case, I am not asserting anything to anyone. For example, I do not believe that Grice's first maxim of quality is in effect when I say something to my companion (who is in on my little scheme). So, while the eavesdropper may *believe* that she has overheard an assertion, she is wrong. And, since there is no assertion, there is no lie.

*IDL** does not appear to be too broad. For example, it correctly rules out statements that are believed to be false, but which are followed with a wink or are part of a play. (In such cases, the speaker does not intend to deceive his audience.) However, this definition is too narrow. As several philosophers have pointed out, an intention to deceive is not a *necessary* condition on lying.[41] Thus, *IDL** incorrectly rules out many things that are lies.

Admittedly, lying very often does involve an intention to deceive. For example, Silvio clearly intended to deceive the jury when he testified that "Tony was home with me at the time of the murder." He wants to deceive them because he wants Tony to be acquitted. Similarly, I intend to deceive my new acquaintance at the party so that she will be impressed with me.

However, there are also cases of lying that do not involve an intention to deceive.[42] For example, suppose that it is a regular citizen (instead of a member of Tony's crew) that has been called as a witness to testify about Tony's whereabouts at the time of the murder. Such a witness might testify falsely about Tony's whereabouts, not because he wants this gangster to be acquitted, but because he fears for his safety if he testifies truthfully.[43] In that case, while he makes a false statement, he does not intend to deceive the jury by doing so. In fact, he may actually hope that the jury will not be deceived, because he will be safer if Tony is convicted and put away. Nevertheless, he is still lying to them.[44]

I do not have to have any *specific* person in mind as my intended audience. For example, a journalist who makes up a news story is lying even if she does not have any idea exactly who her readers will be (cf. Mahon, *op. cit.*, section 1.3). And I may not even have to make the statement to someone that actually exists. For example, a homeowner who is awakened by a suspicious noise and shouts "I have a rifle" may be lying even if there is no burglar (but contrast Chisholm and Feehan, *op. cit.*, pp. 157–58).

[41] See, for example, Siegler, *op. cit.*, p. 129; Shibles, *op. cit.*, p. 102; Carson, *op. cit.*, p. 289; Sorensen, *op. cit.* Everybody does agree that, in order to be a lie, your stating that *p* must be an *intentional action.* For example, when the Jedi Master, Obi-Wan Kenobi, caused an Imperial Stormtrooper to say, "These aren't the droids we're looking for," the Stormtrooper was not lying even if he believed that this statement was false. If someone has taken control of your vocal cords, you are not lying no matter what other conditions might hold. But intending to state that *p* is not the same as *intending to deceive* by stating *p*.

[42] The following example is inspired by Carson (*op. cit.*, pp. 289–90).

[43] Silvio might also fear for his safety if he testifies truthfully. But we can assume that he wants Tony to be acquitted as well.

[44] It might be suggested that this example is somewhat unrealistic. That is, even a regular citizen will probably intend to deceive the jury because succeeding in deceiving them is the only way that he can be sure to escape the wrath of Tony (and his crew). But I will give a more definitive counterexample in just a moment.

While the witness does not intend to deceive the jury, he does at least *foresee* that they are likely to be deceived. This suggests a weaker condition on lying.[45] This condition would also rule out statements that are believed to be false, but that are followed with a wink or that are part of a play. In those cases, the speaker does not expect that his audience might be deceived.

You *lie* to *X* if and only if: (*FDL*)

(1) You state that *p* to *X*.
(2) You believe that *p* is false.
(3) You believe that, by making this statement, *X* may be deceived with respect to *p*.

But not only does a liar not have to have an intention to deceive, a liar does not even have to foresee that someone might be deceived. We can modify the witness example to show this. Suppose that the witness knows that the jury has already seen a videotape of Tony committing the murder.[46] In that case, the witness will have no expectation that the jury will believe his statement about Tony's whereabouts at the time of the murder. Nevertheless, he is still lying to them. This is what is sometimes referred to as a "bald-faced lie" (cf. Sorensen, *op. cit.*).[47]

---

[45] Philosophers sometimes make a distinction between (a) what a person intends to do by performing an action and (b) what a person foresees as a likely "side effect" of performing that action (cf. Carson, *op. cit.*, p. 291; Knobe, *op. cit.*, pp. 208–09). It may be that this distinction does not matter for whether someone is lying.

[46] Cf. Carson, *op. cit.*, p. 289. If such a videotape exists, the witness knows that he is likely to be charged with perjury (that is, lying under oath). But he may very well be more afraid of Tony than he is of a perjury charge.

[47] Tim Kenyon refers to such statements as "cynical assertions." See his "Cynical Assertion: Convention, Pragmatics, and Saying 'Uncle'," *American Philosophical Quarterly*, XL (2003): 241–48. He does not count such statements as lies, but only because he assumes that the standard philosophical definition of lying (*IDL**) is correct. However, there is empirical evidence that most people are disposed to count such statements as lies. Psychologists at the University of Oregon (Taylor and colleagues) wanted to see whether young children could distinguish between lying and pretending. For this study, the psychologists wrote several brief stories where the protagonist lies. In order to create pretending stories, they simply modified the lying stories to remove the intent to deceive. For example, they changed the story so that everyone in the story was aware that the protagonist was not telling the truth. In the experiment, the children correctly identified all of the lying stories as involving lying. However, the children "incorrectly" identified many of the pretending stories as involving lying. When they repeated the experiment using adults instead of children, the psychologists got exactly the same results. Since the "pretending stories" did not have the prototypical features of pretending (such as taking on a role to have fun), the psychologists had arguably created stories that involve bald-faced lying rather than pretending. When the psychologists subsequently went on to create pretending stories that did have such features, the children and the adults were perfect at distinguishing lying from pretending.

Thomas Carson (*op. cit.*, p. 290) has an even better example of this sort. A student has been accused of plagiarism and the student knows that the dean knows that he did it. But the student also knows (based on the dean's reputation) that he will not be punished unless he confesses. So, when the student is called into the dean's office, he denies having plagiarized. Although the student does not expect the dean to be deceived, he is pretty clearly lying to the dean.[48]

These cases show that lying is not always about deception.[49] We typically lie in order to deceive other people. And we want these other people to be deceived because it serves our purposes in some way. For example, my new acquaintance will be more impressed with me (or so I believe) if she is deceived about my occupation. As the witness and plagiarist cases show, however, lies can sometimes serve useful purposes even when they do not deceive (cf. Carson, *op. cit.*, p. 295; Sorensen, *op. cit.*, p. 262).

So, intending to deceive is not a necessary condition on lying. Thus, we need some other way to rule out statements that are followed with a wink or that are part of a play.

VI. ASSERTING

A number of philosophers (for example, Chisholm and Feehan, *op. cit.*, p. 152; Adler, *op. cit.*, p. 435; Watson, *op. cit.*, p. 72; Sorensen, *op. cit.*, p. 256) agree with me that lying is simply asserting what you believe to be false. Asserting has a normative component that goes beyond merely making a statement.[50] And it is this normative component that will rule out statements that are followed with a wink or that are part of a play.

---

[48] Kant, "Ethical Duties towards Others," p. 227 actually describes a similar case where "my enemy takes me by the throat and asks where I keep my money." In that case, "the thief knows full well that I will not, if I can help it, tell him the truth." However, Kant concludes that "my untruth is not a lie because the thief ... has no right to demand [the truth] of me." In addition, Kant, *Foundations of the Metaphysics of Morals*, p. 40, suggests that, if an assertion is known to be false, we "would only laugh at any such assertion as vain pretense." In other words, there would be no point in lying in such a case. But there is a point to the student lying about having plagiarized even if he does not expect to deceive anyone. Moreover, the plagiarist case is a counterexample to *IDL** as well as *FDL*. The student does not expect to change the dean's mind about his guilt. And it is not clear that you can *intend* to do something that you do not expect to succeed at (cf. Alfred R. Mele, "Intention, Belief, and Intentional Action," *American Philosophical Quarterly*, XXVI (1989): 19–30, here pp. 19–20). But even if the student could intend to change the dean's mind, he still might not intend to do so. The student knows that he can get what he wants (namely, getting away with plagiarism) simply by asserting his innocence. So, he may have no interest at all in actually convincing the dean of his innocence.

[49] Even if the speaker does not have to intend to deceive, it might be suggested that someone involved must intend that someone be deceived (cf. Siegler, *op. cit.*, pp. 129–30). For example, Tony presumably intends that the jury be deceived about his whereabouts. But it is not clear that anyone involved in the plagiarist case intends that anyone be deceived.

While a number of philosophers agree with me that *AL* is the correct definition of lying, only a few (for example, Chisholm and Feehan, *op. cit*; Carson, *op. cit.*) have tried to specify precisely what the normative component of assertion is.[51] In this section, I will argue that these accounts of assertion (unlike my own) yield the wrong results when they are plugged into *AL*.

VI.1. Chisholm and Feehan on Asserting. Roderick Chisholm and Thomas Feehan (*op. cit.*, p. 152) think that whether you have asserted something depends on exactly what you believe. In particular, you *assert* something when (a) you say something and (b) you believe that you are in a situation where your audience is justified in believing (i) that you believe what you say and (ii) that you intend that they believe that you believe what you say. Plugging this account of assertion into *AL*, we get Chisholm and Feehan's definition of lying.[52]

You *lie* to *X* if and only if: (*CFL*)

(1) You state that *p* to *X*.
(2) You believe that, by making this statement, *X* becomes justified in believing that you believe *p*.
(3) You believe that, by making this statement, *X* becomes justified in believing that you intend *X* to believe that you believe *p*.
(4) You believe that *p* is false.

This definition of lying correctly rules out statements that are followed with a wink or that are part of a play (cf. Chisholm and Feehan, *op. cit.*, p. 152). In these cases, the speaker does not believe that his audience is justified in believing that he believes what he says.

However, this definition of lying is clearly too narrow. Essentially, under this definition, you are only lying if you expect that you will be successful in deceiving someone about what you believe.[53] But when I seriously say to my new acquaintance, "I am the Prince of Denmark," I do not necessarily think that she is justified in believing that I am the Prince of Denmark or that I believe that I am. She may very well be suspicious of strange men at fancy parties who claim to be royalty. Even so, I am clearly lying to her when I make this statement.[54]

---

[50] Merely stating something has a normative component. For example, we have to obey many semantic and syntactic rules in order to communicate with each other (cf. Williams, *op. cit.*, p. 86). But asserting something has a normative component that involves truth (and/or justification) as well as meaning (cf. Chisholm and Feehan, *op. cit.*, p. 151).

[51] Sorensen (*op. cit.*, pp. 255–56) offers another account of assertion. He claims that an assertion must have "narrow plausibility." In other words, "someone who only had access to the assertion might believe it." (Assertions need not have "wide plausibility." Bald-faced lies, for example, are assertions.) But it is not clear to me why statements that are followed with a wink or that are part of a play do not have such narrow plausibility. In other words, it is not clear to me why the speaker in such cases is only *pretending to assert* rather than *actually asserting* on Sorensen's account.

[52] O'Neill, *op. cit.*, has recently proposed a formal definition of lying that is very similar in spirit to Chisholm and Feehan's definition.

However, the Chisholm-Feehan definition can easily be modified to avoid this problem. Instead of requiring that the listeners be justified in believing that the speaker believes what she says, we can simply require that the listeners be *given a reason to believe* that the speaker believes what she says.

You *lie* to *X* if and only if: (*CFL**)

(1) You state that *p* to *X*.
(2) You believe that, by making this statement, *X* has been given a reason to believe that you believe *p*.
(3) You believe that, by making this statement, *X* has been given a reason to believe that you intend *X* to believe that you believe *p*.
(4) You believe that *p* is false.

The fact that someone says something is usually some evidence that he believes it. For example, when I seriously say to my new acquaintance, "I am the Prince of Denmark," I *do* think that she has been given a reason to believe that I am the Prince of Denmark and that I believe that I am. This evidence that a speaker believes a particular proposition may sometimes be outweighed by other evidence that he does not believe it. But it is still evidence that he believes it.[55]

In some circumstances, however, the fact that someone says something is no evidence at all that he believes it. For example, the fact that an actor on stage says, "I am the Prince of Denmark," gives us no reason to believe that he really believes that he is the Prince of Denmark.

---

[53] More precisely, you expect that, if you end up being unsuccessful in deceiving someone about what you believe, it will be the fault of your listeners, who fail to believe what they are justified in believing.

[54] In fact, I am still lying to her even if I do not expect her to believe me. Huss (*op. cit.*, pp. 128–29) concludes that bluffing is not the same as lying on the grounds that you can bluff someone even when she suspects that you might be trying to deceive her. However, you can also lie to someone even when she suspects that you might be trying to deceive her. In such a case, you do not expect your listener to simply believe what you say. Even so, you can still intend to deceive her with your lie. In particular, you may hope to at least increase her degree of belief in a false proposition (cf. Daya Krishna, "'Lying' and the Compleat Robot," *British Journal for the Philosophy of Science*, XII (1961): 146–49, here p. 147; Chisholm and Feehan, *op. cit.*, p. 145).

[55] According to Richard Foley, we cannot simply equate *evidence* with *reasons for belief*. See his "Evidence and Reasons for Belief," *Analysis*, LI (1991): 98–102. But the few exceptions to this equivalence can be safely ignored for purposes of this paper.

In addition, there are *lies* that give us no reason to believe that the liar believes what she says. For example, the student knows that the dean has no reason to believe that the student believes what he says. Also, the student knows that the dean has no reason to believe that the student intends the dean to believe that the student believes what he says. Everybody involved in this situation knows that everybody is just going through the motions. Nevertheless, the student has clearly *asserted* his innocence and, thus, lied to the dean. Similarly, when someone who is known to be an inveterate liar makes a statement, there is no reason for anyone to believe that she believes that the statement is true. So, if she knows that she is known to be an inveterate liar, the conditions of *CFL** will not be satisfied. But presumably, someone who is known to be an inveterate liar can still lie.[56] Thus, *CFL** is still too narrow.[57] As Carson puts it, "Chisholm and Feehan's definition has the very odd and unacceptable result that a person who is notoriously dishonest couldn't tell lies to those he knows distrust him" (*op. cit.*, p. 292).

VI.2. Carson on Warranting the Truth. Carson does not think that lying is asserting something that you believe to be false. In fact, Carson (*op. cit.*, pp. 299–301) explicitly argues against several philosophers who hold this view of lying. Instead, according to Carson, (setting aside a few complications) you lie if you *warrant the truth* of something that you believe to be false (*op. cit.*, p. 298).[58] And one warrants the truth of a statement when one makes the statement in a context where "one promises or guarantees, either explicitly or implicitly, that what one says is true" (Carson, *op. cit.*, p. 294). Furthermore, it is important to note that whether one has warranted the truth of a statement is independent of what one intends or *believes* (cf. Carson, *op. cit.*, pp. 296–98).

---

[56] By contrast, a *pathological liar* (as opposed to someone who merely has a long reputation for lying) may not be a liar at all (cf. Ekman, *op. cit.*, p. 334). As with the Imperial Stormtrooper, it is not clear that his stating that p is an intentional action.

[57] It would not help to weaken the condition so that it is only required that the listeners be *given what they take to be a reason to believe* that the speaker believes what she says. In the plagiarist case and the inveterate liar case, the statement is not actually evidence that the liar believes what she says *and* it is not taken by the listeners to be evidence.

[58] One of these small complications will be discussed below in the text. The other two are that, according to Carson, (a) your statement actually has to be false and (b) you only have to fail to believe that it is true. As noted above in section IV, it does not seem that a lie actually has to be false. Also, if a speaker just fails to believe that his statement is true, it sounds to me more like *bullshitting* than lying. According to Frankfurt, *op. cit.*, bullshitting requires a *lack of concern* for the truth rather than just ignorance of the truth. However, *asserting* something when you do not know whether it is true does indicate a lack of concern for the truth. (Of course, as Frankfurt (*op. cit.*, p. 341) points out, bullshitting and lying are not necessarily mutually exclusive categories.) But in any event, my definition can easily be modified to incorporate both of these complications. As I discuss in the text, my main objections to Carson's definition lie elsewhere.

You *lie* to *X* if and only if: (*CL*)

(1) You state that *p* to *X*.
(2) You warrant the truth of *p* by making this statement.
(3) You believe that *p* is false.

Despite his rejection of *AL*, however, Carson's definition of lying is equivalent to saying that (a) lying is asserting something that you believe to be false and (b) warranting the truth is the normative component of assertion (cf. Carson, *op. cit.*, p. 300). In fact, several philosophers have characterized the normative component of assertion in exactly this way.[59] According to these philosophers, whenever we make an assertion, we promise or guarantee to our audience that what we say is true. In addition, our audience has the right to rebuke us if our claim turns out to be false. Thus, when we assert *p*, we take on a certain responsibility with respect to *p*. We do not take on the responsibility of actually making *p* true (cf. Carson, *op. cit.*, p. 294). But we do make a "commitment to the defensibility of *p*" (Watson, *op. cit.*, p. 68).

CL gets right the cases that we have considered so far. For example, it rules out statements that are followed with a wink or that are part of a play. In these cases, the speaker does not offer a guarantee that what she says is true. But it rules in the statements made by the plagiarist and the inveterate liar. In these cases, the speaker clearly warrants the truth of what she says.

As Carson (*op. cit.*, pp. 297–98) himself recognizes, however, *CL* is too broad as it stands. For example, suppose that a politician has been asked to give a serious speech at one banquet and a humorous speech at another banquet. Further suppose, however, that the politician gets his dates mixed up and accidentally delivers the humorous speech to the audience expecting a serious speech. The humorous speech includes a joke about the President having "broken wind" at a meeting with foreign dignitaries (an event which did not actually occur). Carson's intuition is that the politician is not lying in this case even though he says something that he believes to be false. However, since this is a context where his audience expects a serious speech, the politician (according Carson's definition) warrants the truth of his statements.

---

[59] See, for example, Charles Sanders Peirce, "Belief and Judgment," in *Collected Papers*, Volume V, Charles Hartshorne and Paul Weiss, eds. (Cambridge: Harvard, 1934), pp. 376–87; Brandom, *op. cit.*; Watson, *op. cit.*

In order to rule out such cases, Carson (*op. cit.*, p. 298) adds an additional condition to his definition. According to his revised definition, the politician is not lying because he *believes* that he is not warranting the truth of his statements.

You *lie* to *X* if and only if: (*CL**)

(1) You state that *p* to *X*.
(2) You warrant the truth of *p* by making this statement.
(3) You do not believe that you are not warranting the truth of *p* by making this statement.
(4) You believe that *p* is false.

However, *CL** is still too *narrow*. This can be shown by considering a simple variation on Carson's own example. Suppose that, when the politician gets his dates mixed up, he accidentally delivers the serious speech to the audience expecting a humorous speech. Given that he believes that his audience expects a serious speech, my strong intuition is that the politician *is* lying in this case if he makes statements that he believes to be false.[60] However, since this is a context where his audience actually expects a humorous speech, the politician does not (according Carson's definition) warrant the truth of his statements.[61] Thus, according to Carson's definition of lying, the politician is not lying.

VI.3. Believing that One is Warranting the Truth. According to Carson's definition of warranting, whether a speaker has warranted the truth of a statement is independent of what the speaker intends or *believes*. However, it seems that his beliefs do matter to whether he is lying.[62] In both versions of the mixed-up politician case, for example, what the politician *believed* that he was doing seemed to be a critical factor in our judgment about whether he was lying. This is right in line with Saint Augustine's claim that "a person is to be judged as lying or not lying according to the intention of his own mind, not according to the truth or falsity of the matter itself."[63] Thus, instead of saying that a speaker has to actually warrant the truth in order to assert something, we might just say that a speaker has to *believe* that he is warranting the truth.

---

[60] We actually do not have to rely on intuitions about cases where a speaker unknowingly fails to warrant the truth of a statement in order to see that warranting the truth of a statement is not necessary for lying. In section VI.3 below, I give a more mundane counterexample to *CL** where a speaker lies despite explicitly telling his audience that he does not warrant the truth of his statement.

[61] The politician *believes* that he is warranting the truth of what he says. But he is wrong (cf. Carson, *op. cit.*, p. 297).

[62] This is certainly suggested by the condition that Carson added to his definition. While some speech acts (such as *saying* something) can be performed without knowing that you are doing it, other speech acts (such as *promising* something) require that you know that you are doing it. See David Owens, "A Simple Theory of Promising," *Philosophical Review*, CXV (2006): 51–77, here p. 54. Lying seems to fall into this second category as well.

You *lie* to *X* if and only if: (*BWL*)

(1) You state that *p* to *X*.
(2) You believe that you warrant the truth of *p* by making this statement.
(3) You believe that *p* is false.

BWL gets both versions of the mixed-up politician case correct. But unfortunately, this definition still rules out some cases of lying that should not be ruled out. For example, suppose that the witness follows up his statement that "Tony was with me at the time of the murder" by saying, "Of course, you know I am really bad with dates and times." This proviso makes it clear that the witness is not *guaranteeing* that Tony was with him at the time of the murder.[64] Thus, he is not (according Carson's definition) warranting the truth of what he says.[65] In addition, he clearly *does not believe* that he is warranting the truth of what he says. Thus, he is not (according to *BWL*) lying. However, if he believes that Tony was not with him at the time of the murder, it still seems pretty clear that he has lied to the jury. The witness is not taking any responsibility for the truth (or defensibility) of his statement about Tony's whereabouts at the time of the murder. (If the jury decides to rely on his statement and it turns out to be false, the responsibility lies with them.) Yet it still seems that he intends to be taken seriously as having expressed what he believes to be the case. In other words, he is still making an assertion about Tony's whereabouts at the time of the murder (cf. Sorensen, *op. cit.*, p. 255). Thus, this case shows that you can be lying (when you say something that you believe to be false) even if you explicitly say that you are not warranting the truth of what you say.[66]

---

[63] Augustine is talking specifically about what the speaker believes about the statement itself. But it seems that the point applies more generally. Although it is ultimately unsuccessful, a virtue of Chisholm and Feehan's definition is that it does cash things out solely in terms of what the speaker believes. Peter Graham, *op. cit.*, makes an analogous point with respect to defining *testimony*.

[64] The witness may be lying about being bad with dates and times (as well as about Tony's whereabouts). But regardless of whether he actually is bad with dates and times, his merely *saying* that he is bad with dates and times means that he is not warranting the truth of his statement about Tony's whereabouts. (Thanks to Kay Mathiesen for this example.)

[65] This case is a counterexample to *CL** as well as *BWL*. The witness is lying even though he is not actually warranting the truth of his statement.

VII. SOME VIRTUES OF MY DEFINITION

My definition correctly handles the counterexamples to the other proposed definitions of lying. The plagiarist, the inveterate liar, the politician giving a serious speech, and the witness who claims to be bad with dates and times all believe that they are in a situation where they should not say what they believe to be false. Thus, according to *BNL*, they are lying when they say something that they believe to be false.

In addition, my definition is not open to the sort of counterexample considered in the previous section. For example, suppose that the witness follows up his statement that "Tony was with me at the time of the murder" by saying, "Of course, I think that it is okay to say things that I believe are false in this sort of situation." Following his statement with this proviso is just like following his statement with a wink (to indicate that he is not to be taken seriously). Thus, this proviso makes it clear that Grice's first maxim of quality is not currently in effect. In addition, he clearly *does not believe* that Grice's first maxim of quality is currently in effect. Thus, he is not (according to *BNL*) lying. But this seems like exactly the right result in this case. He cannot explicitly say that he does not take himself to be subject to Grice's first maxim of quality and still be making an assertion. In fact, he is essentially saying with this proviso that he is not making an assertion. Thus, he does not seem to be lying when he says something that he believes to be false about Tony's whereabouts at the time of the murder.

Another virtue of my definition is that it provides a straightforward analysis of Augustine's (*op. cit.*, p. 57) example of an *altruistic lie*.[67] A man tells his friend that there are no bandits on a certain road even though he believes that there are. He does this because he knows that his friend does not trust him and will conclude from his statement that there *are* bandits on the road.[68] This person is lying under my definition (*BNL*), because he says something that he believes to be false and he believes that Grice's first maxim of quality is in effect. But, surprisingly enough, he is not lying under the standard philosophical definition of lying (*IDL**). This person does not intend to deceive his friend about there being no bandits on the road.[69] He wants his friend to believe correctly that there *are* bandits on the road. In fact, this person does not even intend to deceive his friend with respect to his believing that there are no bandits on the road.[70]

---

[66] There is a counterexample to Watson's (*op. cit.*, pp. 67–70) account of the norm of assertion along the same lines. In that case, the relevant proviso is "Of course, I am not committed to the defensibility of what I just said." Admittedly, this counterexample is not quite as felicitous sounding as my counterexample to *BWL*. But it still seems to work. And my conjecture is that any account of the norm of assertion other than my own can be handled in the same sort of way.

[67] The scientist that fakes her data in order to trick people into believing a theory that she believes to be true is another example of an altruistic lie (cf. O'Brien, *op. cit.*, p. 228). This example counts as a lie on both my definition and the standard philosophical definition.

VIII. SOME OBJECTIONS TO MY DEFINITION

I am a big fan of the television show *Homicide.* Regularly on this show, the homicide detectives put suspects in "the box" and try to get them to confess. A standard technique for achieving this goal is to lie to the suspect.[71] For example, the suspect is told that his fingerprints have been found on the murder weapon, that his DNA has been found at the crime scene, or that his partner has just given him up.

This sort of case might seem to be a counterexample to my definition of lying. The cops are pretty clearly lying to the suspect. But there does not seem to be a norm in effect that the cops should not say what they believe to be false (quite the contrary, in fact). If that were true, they would not be lying according to my definition. Thus, my definition would be too narrow. However, I contend that the norm *is* in effect in "the box" and that the cops believe that it is.[72]

The cops believe that they are in a situation where they should not say things that they believe to be false. It is just that other interests of the cops (namely, getting the suspect to confess) override this norm (cf. Simon, *op. cit.*, pp. 211–12).[73] In other words, the norm is in effect and the cops have violated it, but they have a good *excuse* for doing so.[74] Similarly, the posted speed limit is still 35 miles per hour even if I decide that it is best to rush my injured friend to the hospital at 70 miles per hour. I have broken the speed limit even though I may have a good excuse for doing so.

---

[68] In the same passage, Augustine also gives an example of a *deceptive truth.* A man truthfully tells his enemy that there are bandits on a certain road. He does this because he knows that his enemy does not trust him and will conclude from his statement that there are no bandits on the road. Although he does intend to deceive his enemy with his statement, this person is not lying on both my definition and the standard philosophical definition.

[69] This person does intend to deceive his friend. In particular, he intends that his friend believe that he intends that his friend believe that there are no bandits on the road. Thus, this case would count as a lie under *IDL.* But, as noted above in section v, *IDL* is too broad as a definition of lying.

[70] Some definitions require an intention to deceive with respect to your believing p in addition to (or instead of) an intention to deceive with respect to p itself (cf. Mahon, *op. cit.*, section 1.6).

[71] Cf. David Simon, *Homicide* (New York: Ivy, 1991), pp. 216–17; Christopher Slobogin, "Deceit, Pretext, and Trickery: Investigative Lies by the Police," *Oregon Law Review*, LXXVI (1997): 775–816, here pp. 785–88.

[72] The cops are clearly making use of the fact that at least the suspect believes that the norm is in effect.

Everyone involved in the interrogation seems to recognize that the norm is in effect. For example, after he learns of the deception, the suspect will often complain that he has been lied to. In addition, while the cops rarely have much sympathy for the suspect, they do seem to recognize that it is reasonable for him to complain. By contrast, the norm is not in effect at all if you are performing a play or if you wink to indicate that you are not to be taken seriously. If a member of the audience complained that she had been lied to, she would have to be crazy and the cast would undoubtedly think that she was crazy.[75]

I am also a big fan of *Star Wars*, and my favorite character is the pilot of the *Millennium Falcon*, Han Solo. In one memorable scene, our heroes use the garbage chute to escape from the detention block of the *Death Star*. Once they land in the garbage, Han Solo sarcastically says, "The garbage chute was a really wonderful idea. What an incredible smell you've discovered!"

---

[73] The norm is *explicitly* in effect for a witness who is testifying under oath, but the witness may have interests (for example, avoiding the wrath of Tony and his crew) that outweigh the force of the norm. A *white lie* is another example where Grice's first maxim of quality is in effect, but is overridden by other interests (for example, not offending someone) that we have. In addition to coming into conflict with other interests that we have, the norm in question ("Do not say what you believe to be false") can come into conflict with other norms. For example, it can come into conflict with the moral norm "One ought to prevent harm to others" (à la W.D. Ross, *The Right and the Good* (New York: Oxford, 1930)) as when the only way to save a friend from a murderer is to lie (cf. Kant, *Critique of Practical Reason*). But my claim is that, regardless of what else is going on, disobeying this norm when you believe that it is in effect is lying.

[74] Norms of conversation can "clash" with each other as well as with other interests that we have (Grice, *op. cit.*, p. 30). For example when a teacher who believes in creationism has to give a lesson on evolution, Grice's first maxim of quality comes into conflict with Grice's second maxim of quality. If the teacher violates the norm against saying what she believes to be false solely in order to obey the norm against saying that for which she lacks adequate evidence, some (for example, Lackey, *op. cit.*, p. 602) might want to say that she is not lying. In order to accommodate that intuition, my definition might be modified to include an exemption for such cases.

[75] There are other contexts in which is okay to violate the norm; but not simply because the speaker has a good excuse. For example, the *Liar's Club* was a television game show on which a panel of celebrity guests passed around some unusual object and gave bizarre explanations of its purpose. The contestants then had to guess which of the celebrity guests was *not* lying to them. The celebrity guests had essentially been given permission to violate the norm in this context. (It is perfectly okay to bluff in poker for the same reason.) Thus, even though they have also been lied to, it would be crazy for contestants on this show to complain about it in the way that it is not crazy for the suspect to complain about it.

This sort of case might also seem to be a counterexample to my definition of lying. Solo is pretty clearly not lying.[76] Just the opposite, in fact. He is trying to communicate something that he believes to be true (namely, that the garbage chute was a really *bad* idea). But he is certainly *saying* something that he believes to be false. In addition, Grice's first maxim of quality is arguably in effect. If that were true, he would be lying according to my definition. Thus, my definition would be too broad. However, I contend that the norm is not in effect with respect to Solo's sarcastic comment and that Solo does not believe that it is.

Solo's comment about the garbage chute is an example of *conversational implicature.*[77] Solo "blatantly fails to fulfill" or "flouts" the first maxim of quality in order to express something other than what he literally says (Grice, *op. cit.*, p. 30). I contend that, by flouting this norm of conversation, Solo turns it off.

The default whenever we engage in linguistic behavior is that the first maxim of quality (as well as the rest of Grice's maxims) is in effect. However, these norms can be turned off or suspended. Basically, you can signal to your audience—in a manner that is collectively accepted by the relevant linguistic community for this purpose—that you will not be obeying the norm of conversation in question.[78] If you give such a signal, other people will not expect you to obey the norm, they will not think that you ought to obey the norm, they will not think that you are subject to rebuke for not obeying the norm, and so on. Winking is an example of a signal that turns off Grice's first maxim of quality in many cultures.

---

[76] Han Solo is not above telling lies (for example, about the speed of the *Millennium Falcon*). But he is not lying here. I would like to thank Daniel Silvermint for this example of conversational implicature. I would also like to thank Marc Moffett for pressing this objection in his excellent comments on this paper at the 2008 Pacific Division Meeting of the American Philosophical Association.

[77] Cf. Grice, *op. cit.*, pp. 24–37; Wayne Davis, "Implicature," *Stanford Encyclopedia of Philosophy* (2005): http://plato.stanford.edu/entries/implicature/.

[78] In other words, there is a *norm* that a particular signal turns off the norm of conversation in question. Of course, even if they do not believe that she intended to send such a signal, it will sometimes be clear to an audience that a speaker is not to be taken seriously. For example, if a witness in court says with a straight face that she is a Giant Panda, the best explanation may be that she is crazy. Now, if she really is crazy, but believes that she is actually a Giant Manatee, then she is clearly lying. But if she is not crazy and is only trying to convince everyone that she is crazy or is just trying to disrupt the proceedings, some (for example, Mahon, *op. cit.*, section 1.6) might want to say that she is not lying. In order to accommodate that intuition, it might be suggested that pretending to be crazy makes it reasonable for one to believe that the norm is not in effect (even though it does not make it common knowledge that the norm is not in effect as winking would).

While a collectively accepted signal is often an overt behavior such as a wink, it can also be quite subtle. In fact, it may just be clear from the context that the speaker does not really mean to assert seriously what he is actually saying. For example, the deadpan humorist Steven Wright is not lying when he says with a completely straight face, "I can levitate birds. No one cares."[79] Similarly, when a speaker flouts Grice's first maxim of quality as Han Solo did, the best explanation of the utterance is that the speaker means to express something other than what he is actually saying.[80]

A norm of conversation does have to be in effect at the time of utterance in order for conversational implicature to work. For example, the fact that Grice's first maxim of quality is in effect (and the fact that Solo is clearly not acting in accordance with it) is part of what allows Princess Leia to figure out what Solo really means to express. But in the case of both winking and flouting, the norm *is* in effect when the utterance begins. The signal essentially has the effect of retroactively turning off the norm.[81] For example, we usually wink *after* we make a statement that we do not want to be taken seriously.[82]

By contrast, a bald-faced liar, such as the plagiarist, gives no such signal. Although he blatantly fails to fulfill the norm, it is still in effect.[83] The best explanation of his utterance is that he means to express exactly what he says. Unlike Solo, he is lying.

IX. DECEPTIVE LYING

While lying does not require an intention to deceive, it must be acknowledged that philosophers are primarily interested in lies that are intended to deceive.[84] For example, Kant is clearly concerned with cases of lying that have the potential to destroy trust. He claimed that it is always wrong to lie because lying whenever it is to your advantage is a self-defeating moral rule. If everyone followed this rule, we could not trust anybody when they spoke and there would be no point in lying.[85] Thus, it is not a problem for Kant if things like the plagiarist case are ruled out by a definition of lying (cf. Sorensen, *op. cit.*, p. 263). There is certainly something morally problematic about what the student does in the plagiarist case.[86] But it is not clear that it destroys (or even decreases) trust. As Sorensen (*op. cit.*, p. 252) puts it, "bald-faced lies do not fool anyone. They are no more a threat to truth telling than sarcastic remarks."

---

[79] It is not just the absurdity of the content that provides the indication. A completely preposterous statement can still be a (bald-faced) lie (cf. Sorensen, *op. cit.*, p. 253).

[80] Davis (*op. cit.*, section 8), for example, suggests that people use *inference to the best explanation* to work out conversational implicatures.

[81] It should also be noted that the signal typically turns off the norm with respect to that one single utterance. However, some might object that, instead of actually turning off the norm, flouting just makes disobeying the norm permissible. In that case, another condition would simply need to be added to my definition of lying. Namely, you must not believe that you have signaled to *X*—in a manner that is collectively accepted by the relevant linguistic community for this purpose—that you mean to express something other than *p*.

[82] Interestingly, it seems that the signal can be sent fairly long after the statement is made and still turn off Grice's maxim with respect to that statement. For example, some people will not give any indication that they are joking until they have you completely convinced that they are serious. But such "leg pulling" does not seem to count as lying.

[83] Grice (*op. cit.*, p. 30) seems to suggest that you can only "quietly and unostentatiously *violate* a maxim." But there is certainly a sense in which the plagiarist "blatantly" violates the first maxim of quality.

In addition, when epistemologists talk about lying, they are principally concerned with *deceptive* testimony that can interfere with people's ability to acquire knowledge. Thus, it may be no problem for epistemologists if things like the plagiarist case are ruled out by a definition of lying. There is nothing epistemically problematic at all about what the student does in the plagiarist case. In particular, it is not going to lead the dean to acquire false beliefs.

Since we are looking for a definition of lying for *philosophical* use (albeit one that comes close to common usage), perhaps our definition should include an "intent to deceive" condition after all. Thus, it might be suggested that the standard philosophical definition (*IDL**) is a perfectly good stipulative definition for our purposes. But although *IDL** would be a simpler definition of *deceptive lying*, I would like to suggest that it would actually be better to add an "intent to deceive" condition to my definition of lying rather than revert to the standard philosophical definition.

We would presumably like a unified account of lying. In particular, deceptive lying should be a proper subset of lying in general. However, *IDL** seems to rule in cases of verbal deception that are not lies at all. For example, an actor on stage is not lying when he recites a line that he happens to believe to be false. But it is possible that the performance is part of an elaborate deception aimed at getting members of the audience to believe that that particular line from the play is actually true. In other words, the actor says something that he believes to be false with the intent to deceive people about that very thing.[87] By contrast, my proposal correctly rules out such cases because Grice's first maxim of quality is not in effect in this context.[88]

---

[84] The very title of Chisholm and Feehan's (*op. cit.*) paper suggests that this is the type of lying that they were specifically interested in.

[85] I do not take any position on whether Kant's argument is correct. (In fact, Elliott Sober argues that, while a world in which everybody lies is unstable, it is not impossible. See his "The Primacy of Truth-Telling and the Evolution of Lying," in *From a Biological Point of View* (New York: Cambridge, 1994), pp. 71–92.) I just point out that the argument requires a type of lying that involves the intent to deceive.

[86] Thus, as Sorensen (*op. cit.*) points out, bald-faced lies clearly have some philosophical interest.

X. CONCLUSION

According to the standard philosophical definition, lying is saying something that you believe to be false *with the intent to deceive*. However, people sometimes "go on the record" with something that they believe to be false even though they have no intent to deceive. In response to such examples of nondeceptive lying, it has been suggested that lying is simply *asserting* something that you believe to be false (where asserting something involves a normative component that goes beyond merely saying something). However, the existing accounts of asserting something give us a definition of lying that still does not capture common usage. In this paper, I have argued that the correct definition of lying is that (a) you say something that you believe to be false and (b) you believe that you are in a situation where the following norm of conversation is in effect: *"Do not say what you believe to be false."* But I do note that philosophers typically *are* interested in lies that are intended to deceive. So, for most philosophical purposes, I propose a definition of *deceptive lying* which simply adds an "intent to deceive" condition to my definition of lying.

DON FALLIS

University of Arizona

---

[87] In order to satisfy the first condition of *IDL**, it would also have to be a line that the actor addresses *to the audience* of the play (for example, in an aside) rather than to another actor. As with this case of the deceptive actor, it is not clear whether an adult telling a tall tale to credulous children is actually lying to them even if he intends to deceive them and, thus, satisfies the conditions of *IDL** (cf. footnote 21 above).

[88] There may be one possible role for the standard philosophical definition. It might turn out to be the case that young children learn how to deceive with words before they learn (implicitly) when the norms of conversation are in effect. If so, we might want to refer to some of their (pre-norm) utterances as lies.

THE JOURNAL OF

PHILOSOPHY

Special Issues Available

Parts and Wholes — December 2006
On Bernard Williams — June 2003
Causation — April 2000
Jürgen Habermas and John Rawls — March 1995

The Dewey Lectures:
Michael Dummett — *Truth and the Past*, 2002
Hilary Putnam — *Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind*, 1994
Donald Davidson — *The Structure and Nature of Truth*, 1990
Amartya Sen — *Well-being, Agency, and Freedom*, 1985
John Rawls — *Kantian Constructivism in Moral Theory*, 1980
Ernest Nagel — *Teleology Revisited*, 1977
W.V. Quine — *Ontological Relativity*, 1968

The Woodbridge Lectures:
John McDowell - *Having the World in View: Sellars, Kant, and Intentionality*, 1997

Subscriptions

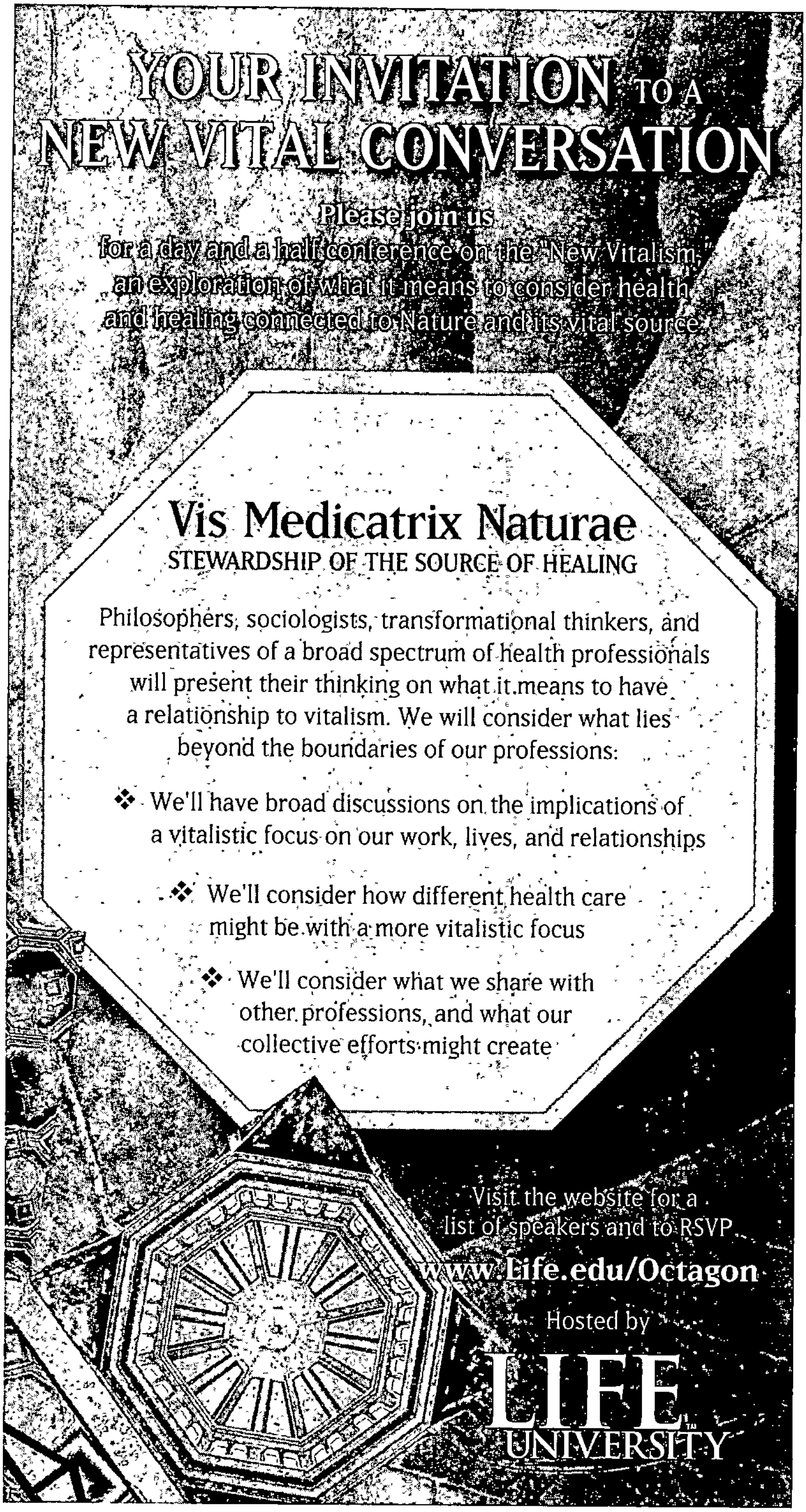
Individuals $45.00
Students $20.00
Libraries and Institutions $100.00
Print and Online for Libraries and Institutions $180.00
Postage outside the United States $15.00

Special Offer

One complimentary issue of any of the above issues with each new subscription.
All issues are $5.00 plus $3.00 for shipping and handling.
Send check or money order to:

The Journal of Philosophy
Columbia University
1150 Amsterdam Avenue, Mail Code 4972
New York, NY 10027

www.journalofphilosophy.org



YOUR INVITATION TO A NEW VITAL CONVERSATION

Please join us

for a day and a half conference on the "New Vitalism,"
an exploration of what it means to consider health
and healing connected to Nature and its vital source

Vis Medicatrix Naturae

STEWARDSHIP OF THE SOURCE OF HEALING

Philosophers, sociologists, transformational thinkers, and representatives of a broad spectrum of health professionals will present their thinking on what it means to have a relationship to vitalism. We will consider what lies beyond the boundaries of our professions:

- We'll have broad discussions on the implications of a vitalistic focus on our work, lives, and relationships
- We'll consider how different health care might be with a more vitalistic focus
- We'll consider what we share with other professions, and what our collective efforts might create

Visit the website for a
list of speakers and to RSVP.

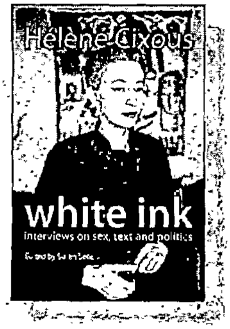
www.Life.edu/Octagon

Hosted by

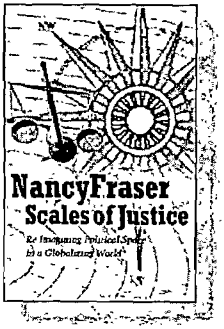
LIFE UNIVERSITY

COLUMBIA

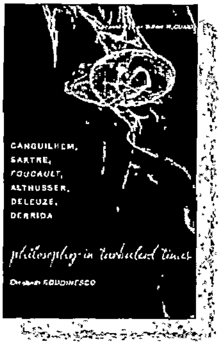
Read book excerpts at www.cup.columbia.edu



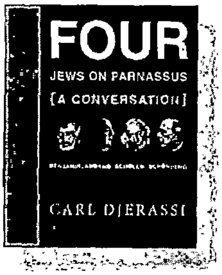
White Ink
Interviews on Sex, Text, and Politics
Hélène Cixous
Edited by Susan Sellers
A rare collection of dialogues that enhance our understanding of the French philosopher's life and work.
paper - $26.50 / cloth - $79.50
European Perspectives: A Series in Social Thought and Cultural Criticism



Scales of Justice
Reimagining Political Space in a Globalizing World
Nancy Fraser
"A groundbreaking collection of essays."—Amelie Rorty, *Harvard University*
cloth - $27.50
New Directions in Critical Theory



Philosophy in Turbulent Times
Canguilhem, Sartre, Foucault, Althusser, Deleuze, Derrida
Elisabeth Roudinesco
Translated by William McCuaig
"Indispensable."
—Gayatri Chakravorty Spivak, *Columbia University*
cloth - $26.50



Four Jews on Parnassus—A Conversation
Benjamin, Adorno, Scholem, Schönberg
Carl Djerassi
Illustrations by Gabriele Seethaler
"Highly innovative, amusing, and insightful."—Jack Zipes, *University of Minnesota*
cloth - $29.50

Horrorism
Naming Contemporary Violence
Adriana Cavarero
Translated by William McCuaig
"Remarkable."
—Arjun Appadurai, *NYU*
cloth - $29.50
New Directions in Critical Theory

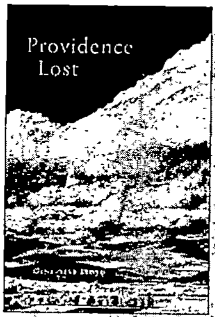
Pragmatism and Social Hope
Deepening Democracy in Global Contexts
Judith M. Green
How America's greatest social and political thinkers point the way toward a more inclusive democratic future.
cloth - $34.50

The Undiscovered Dewey
Religion, Morality, and the Ethos of Democracy
Melvin L. Rogers
"Controversial and interesting."
—Eric MacGilvray, *Ohio State University*
cloth - $45.00

NOW IN PAPER

Freedom and Neurobiology
Reflections on Free Will, Language, and Political Power
John R. Searle
paper - $16.95
Columbia Themes in Philosophy

Geneses, Genealogies, Genres, and Genius
The Secrets of the Archive
Jacques Derrida
Translated by Beverley Bie Brahic
paper - $18.50
European Perspectives

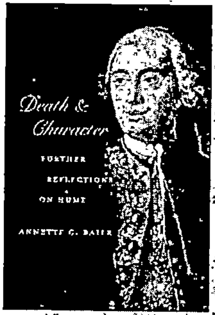


Providence Lost

GENEVIEVE LLOYD

Providence still exerts a powerful influence on our thought and in our lives; and understanding how can help us clarify the functioning—or, increasingly, disfunctioning—of concepts of freedom and autonomy that define our modernity.

new in cloth



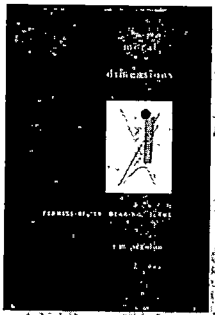
Death and Character

Further Reflections on Hume

ANNETTE C. BAIER

Ranging widely in Hume's works, Baier considers his views on character, desirable character traits, his treatment of historical characters, and his own character as shown by changes in his writings, especially his repudiation of the celebrated *A Treatise on Human Nature.*

new in cloth



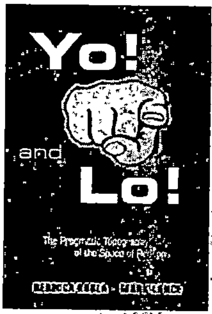
Moral Dimensions

Permissibility, Meaning, Blame

T. M. SCANLON

An action's meaning depends on the agent's reasons for performing it. Blame, Scanlon argues, is a response to the meaning of an action rather than its permissibility. This analysis leads to a novel account of the conditions of moral responsibility and to important conclusions about the ethics of blame.

Belknap Press / new in cloth



'Yo!' and 'Lo!'

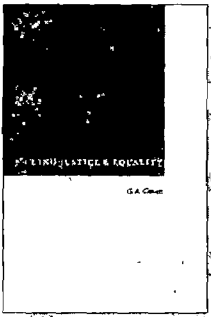
The Pragmatic Topography of the Space of Reasons

REBECCA KUKLA and MARK LANCE

Entering into conversation with the work of Anglo-American philosophers such as Wilfrid Sellars, Robert Brandom, and John McDowell, and Continental philosophers including Heidegger and Althusser, *'Yo!' and 'Lo!'* offers solutions (or dissolutions) to long-standing philosophical problems.

new in cloth

HARVARD
UNIVERSITY
PRESS
www.hup.harvard.edu

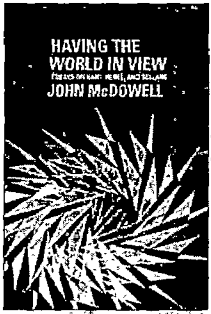


Rescuing Justice and Equality

G. A. COHEN

"With his characteristic brilliance and philosophical depth, G. A. Cohen aims to defend a more pure equality and justice by retrieving them from Rawlsian liberalism. The result is a bracing challenge to contemporary complacency about inequality."
—Joshua Cohen, Stanford University

new in cloth



Having the World in View

Essays on Kant, Hegel, and Sellars

JOHN McDOWELL

Having the World in View will be a decisive further step toward healing the divisions in contemporary philosophy, by showing how central methods of the two traditions remain deeply entangled and by revealing how philosophers in both camps might still learn from each other.

new in cloth

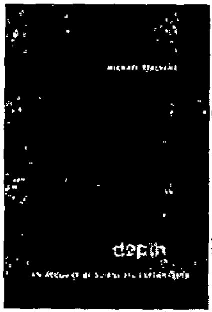
The Engaged Intellect

Philosophical Essays

JOHN McDOWELL

McDowell shows how progress is to be achieved by preserving what is most attractive in the views of those he is in conversation with, while whittling away their weaknesses. As he practices this method, what emerges through the volume is the unity of McDowell's own views.

new in cloth



Depth

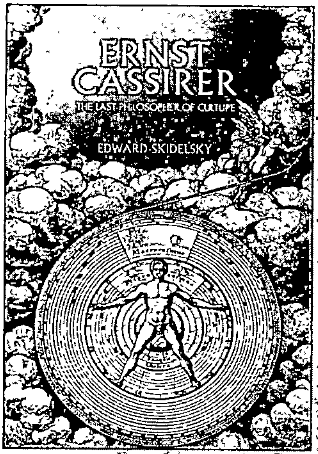
An Account of Scientific Explanation

MICHAEL STREVENS

What does it mean for scientists to truly understand, rather than to merely describe, how the world works? Michael Strevens proposes a novel theory of scientific explanation and understanding that overhauls and augments the familiar causal approach to explanation.

new in cloth

HARVARD
UNIVERSITY
PRESS
www.hup.harvard.edu



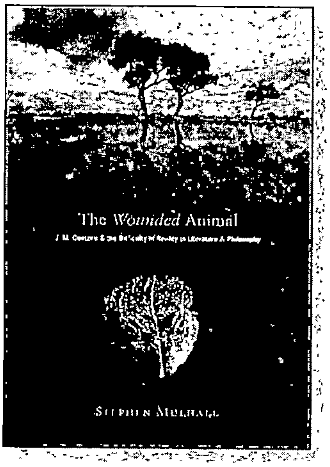
Ernst Cassirer

The Last Philosopher of Culture

Edward Skidelsky

"A major accomplishment and an excellent introduction to Cassirer's philosophy. Most importantly, Skidelsky shows that Cassirer did not simply align himself with the tradition of German political thought reaching from Humboldt and Goethe to the later Thomas Mann, but that his entire philosophy was actually subservient to the project of rethinking the basis of political liberalism. Skidelsky overturns the widely accepted view that Cassirer was an apolitical epistemologist."
—John Michael Krois, author of *Cassirer: Symbolic Forms and History*

Cloth $35.00 978-0-691-13134-4



The Wounded Animal

J. M. Coetzee and the Difficulty of Reality in Literature and Philosophy

Stephen Mulhall

"One of the most suggestive discussions of the relations between philosophy and literature that I have ever read, *The Wounded Animal* is studded with striking insights and penetrating questions."
—Tzachi Zamir, author of *Double Vision: Moral Philosophy and Shakespearean Drama*

Paper $26.95 978-0-691-13737-7
Cloth $75.00 978-0-691-13736-0

PRINCETON
UNIVERSITY
PRESS

800.777.4726
press.princeton.edu

1-7-09
UNIVERSITY
OF
CALCUTTA

THE JOURNAL OF PHILOSOPHY

VOLUME CVI, NUMBER 2
FEBRUARY 2009

page

57 *The "Structure" of Physics: A Case Study* Jill North
89 *Probability in Deterministic Physics* J.T. Ismael
109 NEW BOOKS: TRANSLATIONS

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Michael J. Mooney; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2009

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2009 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 2, FEBRUARY 2009

UNIVERSITY OF CALCUTTA
ALIPORE CAMPUS

THE "STRUCTURE" OF PHYSICS: A CASE STUDY*

We are used to talking about the "structure" posited by a given theory of physics. We say that relativity is a theory *about* space-time structure. Special relativity posits one space-time structure; different models of general relativity posit different space-time structures. We also talk of the "existence" of these structures. Special relativity says that the world's space-time structure is Minkowskian: it posits that this space-time structure *exists*.

Understanding structure in this sense seems important for understanding what physics is telling us about the world. But it is not immediately obvious just what this structure is, or what we mean by the existence of one structure, rather than another.

The idea of *mathematical* structure is relatively straightforward. There is geometric structure, topological structure, algebraic structure, and so forth. Mathematical structure tells us how abstract mathematical objects fit together to form different types of mathematical spaces. Insofar as we understand mathematical objects, we can understand mathematical structure. Of course, what to say about the nature of mathematical objects is not easy. But there seems to be no further problem for understanding mathematical structure.

Modern theories of physics are formulated in terms of these mathematical structures. In order to understand "structure" as used in physics,

---

*For comments and discussion, I am extremely grateful to David Albert, Frank Arntzenius, Gordon Belot, Josh Brown, Adam Elga, Branden Fitelson, Peter Forrest, Hans Halvorson, Oliver Davis Johns, James Ladyman, David Malament, Oliver Pooley, Brad Skow, Ted Sider, Rich Thomason, Jason Turner, Dmitri Tymoczko, the philosophy faculty at Yale University, audience members at The University of Michigan in fall 2006, and in 2007 at the Pacific Division American Philosophical Association meeting, the Joint Session of the Aristotelian Society and Mind Association, and the Bellingham Summer Philosophy Conference.

0022-362X/09/0602/57–88

then, it seems we must simply look at the structure of the mathematics that is used to state the physics.

But it is not that simple. Physics is supposed to be telling us about the nature of the *world.* If our physical theories are formulated in mathematical language, using mathematical objects, then this mathematics is somehow telling us about the physical make-up of the world. What is the relation between these abstract mathematical objects and the physical objects of our experience? What is the relation between the structure of a mathematical space in which we formulate a theory, and the structure of physical space according to that theory? When we infer the existence of a particular kind of structure—the space-time structure of relativity, some other structure posited by a different theory—are we saying something about the world that theory describes, over and above the mathematics needed to formulate the theory?

These questions raise lots of big and interesting philosophical issues—about the existence of mathematical objects, the use of models in science, underdetermination theses, and more. Though relevant, these are not the questions I wish to focus on here.

I want to home in on a different issue. One of the puzzling things about the "structure" of physics is that there can be different mathematical formulations of a given physical theory—and not just gerrymandered ones with dangling bits of mathematics tacked on. Examples of this are ready at hand: Heisenberg's and Schrödinger's formulations of quantum mechanics, Lagrangian and Hamiltonian versions of classical mechanics.

Different mathematical formulations mean different mathematical structures. If there is more than one such formulation, then what can we infer from the theory about the structure of the world? Should we say that different mathematical formulations posit different structures, or are they simply different descriptions of the same underlying structure? When is one theory a mere notational variant of another, and when does it count as a distinct theory, with its own account of what the world's structure is like?

The question of what to infer about the structure of the world from the structure of the mathematics used to state the physics is already difficult. Cases of equivalent mathematical formulations make it all the more so. I want to consider these questions with a specific example in mind. There are many to choose from. I wish to minimize technicality and limit the discussion to classical mechanics, and in particular, the Lagrangian and Hamiltonian formulations. I will suggest some general conclusions, but ultimately this really will be a case study. It remains to work through other cases to see if similar considerations apply.

I. WHAT IS *STRUCTURE*?

When, and why, do we infer the existence of one structure rather than another? When, and why, do we make the jump from the structure of a mathematical space in which we formulate a theory, to the structure of the physical world according to that theory?

These questions require that we get to the bottom of what this "structure" talk is all about. What is this thing we call 'structure', such that two theories can posit different ones?

Start with the structure of a mathematical space; we will work up from there to the structure of physical space. A mathematical space can be endowed with a certain structure, say, a geometric structure. (We can think of a mathematical space as, roughly, a set of points with certain mathematical objects defined on it; these objects give the structure of the space.) We describe such a space in mathematical terms, ascribing to it various mathematical properties. But there is an important difference in the kinds of mathematical features we ascribe to such a space. There is a difference between the intrinsic features of the space, and the features it has merely because of the way we choose to describe it.[1]

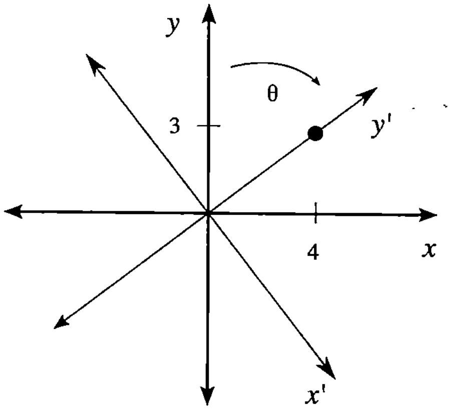
Take a Euclidean plane. In order to describe the different locations in the plane, we lay down a coordinate system on it. This allows us to associate numerical values with the points in the plane. There are many ways we could do this. We might use a rectangular coordinate system. We might use a polar coordinate system. We could use some other, non-rectilinear coordinate system. So long as the coordinates meet certain general constraints (for example, they must vary continuously[2]), we can use any set of coordinates we like to describe the locations in the plane.

The points will get different numerical descriptions depending on which coordinate system we choose. Take a particular point in the plane. Choose a Cartesian coordinate system, and the point will get the coordinates (x, y). Rotate coordinate axes through an angle θ about the origin, and the point will have different coordinates (x', y'). (See Figure 1.) But the *point itself* does not change. Only its *description* changes.

---

[1] For further discussion on the difference between the coordinate-dependent features of a space, on the one hand, and the intrinsic structure of the space, on the other, see Tim Maudlin, *Quantum Non-Locality and Relativity* (Massachusetts: Blackwell, 2002, 2nd ed.), chapter 2, and "Relativity," in Donald M. Borchert, ed., *Encyclopedia of Philosophy*, Volume 8 (Detroit: Thompson Gale, 2006, 2nd ed.), pp. 345–57.

[2] More precisely, the coordinate system within each chart covering the manifold must satisfy the continuity requirement. We could use weirder, non-numerical labels, though that will not help us much when it comes to the physics, where we need differentiable functions on these spaces. I will assume that we are interested in the above sorts of coordinates.



$$(x, y) = (4, 3)$$

$$(x', y') = (0, 5)$$

Figure 1: Points

Take a slightly more complicated object, a vector. Think of a vector as an arrow, a quantity defined by both a magnitude and a direction. Choose a coordinate system, and we can describe the vector in terms of its components. A component is the part of the vector along one of the coordinate axes; the vector is equal to the sum of its components. In the *x*-*y* coordinate system we started out with above, a vector **v** has components v_x and v_y. Rotate coordinate axes, and the vector now has components v'_x and v'_y. (See Figure 2.) The components are different in the new coordinate system. But the vector itself, the arrow with this length and direction, does not change. Only its description changes.

Objects like points and vectors are coordinate-independent, geometric objects: they exist independently of any coordinate system. Change from one coordinate system to another, and their coordinate-

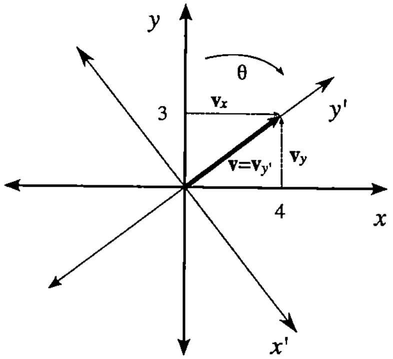


Figure 2: Vectors

dependent descriptions will change; the objects themselves do not. We call these *invariant* objects, since they are invariant under, or unaffected by, changes in coordinates.

This difference between coordinate-independent objects, on the one hand, and their coordinate-dependent descriptions, on the other, is starting to bring out the idea of *structure*. Intuitively, there is a structure to these objects, some intrinsic nature that is unaffected by coordinate changes.

Apply this to the Euclidean plane. We said that there are different coordinates we can use for the plane. The reason is that, like points and vectors, the plane itself has a structure that is independent of coordinates. How do we know this? Because no matter which coordinate system we choose, some of the plane's features always remain the same. No matter what the coordinate system, the distance between any two points in the plane is unchanged. This distance measure is a coordinate-independent feature of the space. It is part of its underlying, geometric structure.

The geometric structure of a mathematical space is given by the geometric objects defined on it. Since geometric objects are invariant under coordinate changes, so too is geometric structure. Geometric structure is given by quantities that remain intact while we alter what are merely arbitrary choices of description. This is what we have in the backs of our minds when we say that we are free to choose different coordinate systems for the plane. We mean that choosing different coordinate systems does not alter the underlying structure. It only alters our *description* of that structure.

Let me briefly say something about the Euclidean distance measure, since you might be wondering about its status as a coordinate-independent object. The familiar Pythagorean theorem, $d = \sqrt{\Delta x^2 + \Delta y^2}$, is not wholly independent of coordinates. Try using nonrectangular coordinates, and this formula will not calculate the distances correctly. That is right, but there is an invariant distance on the plane. This is not given by an algebraic expression like the Pythagorean theorem. It is given by a geometric object called a *tensor*. A tensor is an abstract geometric object that is invariant under coordinate changes, just as points and vectors are. As with points and vectors, tensors get different component descriptions in different coordinate systems. In a rectangular coordinate system, the Pythagorean theorem gives the components of the metric tensor. In a different kind of coordinate system, it does not. But the Euclidean distance between any two points, this number (in a given unit), or scalar, is coordinate-independent. The metric tensor defined on the plane is an invariant object that lets us calculate distances using any allowable coordinate system. Think of it as a generalization of the Pythagorean theorem, legitimate for any set of curvilinear coordinates we can use on the plane.

In general, for any mathematical space, there will be some quantities that all allowable coordinate systems agree on. (If there are no such quantities, then we have a structureless set of points. Such a space would not have enough structure to do physics: it does not even have the topological structure needed to define continuity and differentiability.) These are features that the space will have no matter which coordinate system we use to describe it. These features are part of the intrinsic, coordinate-independent *structure* of the space.

Since quantities that stay put under coordinate changes correspond to intrinsic, structural features of a space, these invariant quantities give us a way of figuring out the structure of a space: choose an allowable coordinate system, change to another one, and see what remains the same when we do this. The fact that coordinate changes preserve Euclidean distances reveals that this metric is part of the plane's underlying structure. Invariant quantities can also indicate a symmetry, or a lack of structure. The fact that coordinate translations and rotations do not disturb the geometry of the plane indicates that this structure is uniform with respect to the different locations and directions in the plane. It indicates that there is no additional structure picking out a preferred point or direction.[3]

Of course, which are the allowable coordinate systems in the first place is determined by that structure. We know which coordinate systems we can use for the plane because we already know what structure it has: the allowable coordinate systems are precisely the ones that preserve the Euclidean distances. What we are seeing is that we also have a way of figuring out the structure of less familiar spaces. Given the allowable coordinate transformations, we can infer the structure from the invariant quantities under those transformations.

Here is where we are so far. The geometric structure of a mathematical space is given by quantities that remain intact under changes in coordinates. There is a difference, then, between the features ascribed to a space by the coordinate system being used, and the intrinsic features of the space itself. There is a difference between (genuine) *structure* and (mere) *description* of that structure.

---

[3] I have been talking in terms of passive transformations, or coordinate changes. Active transformations leave the coordinates alone, acting directly on the coordinate-independent objects to yield changes in their coordinate-dependent descriptions. Both kinds of transformation capture the idea that there *is* some underlying geometric structure of which we can have different, equally good descriptions. Thinking in terms of passive transformations is more intuitive for our purposes, but either will do.

That difference is important in physics, too. In Newtonian mechanics, any reference frame or coordinate system moving with a constant (nonaccelerating) velocity is one in which the laws hold. As far as this theory is concerned, any one of these frames is as good as any other for describing a given system. Choosing one frame rather than another is just an arbitrary choice in description. That is why a frame-dependent quantity like velocity is not a fundamental, objective feature of a Newtonian world. There is no "absolute velocity" on this theory. An object's velocity depends on the choice of frame, any one of which is equally legitimate.

In special relativity, the laws hold in any Lorentz frame. Since the time elapse between events depends on the choice of frame—simultaneity is "relative" to a frame—temporal distances are not in the fundamental structure of the world, according to this theory. They are "merely" frame-dependent quantities.

In physics, the frame-dependent quantities, the quantities whose values depend on the particular choice of frame, are taken to be nonfundamental; they are like the coordinate-dependent features of the plane. Frame-*in*dependent quantities, on the other hand, the quantities which are the same in all allowable reference frames, *do* correspond to fundamental, objective features of the world. In special relativity, all Lorentz frames agree on the space-time interval between two events, just as in the Euclidean plane, all coordinate systems agree on the spatial distance between two points. The space-time interval *is* a fundamental, objective feature of the world, according to the theory of special relativity. It gives the geometry of its space-time, in the same way that the Euclidean metric gives the geometry of the plane.

That is why modern theories of physics are typically formulated in terms of abstract, invariant, geometric objects, rather than numerical, coordinate-dependent, algebraic ones. Objective features of the world are best represented by quantities that do not depend on our arbitrary choices of description—on choices the *theory itself* takes to be arbitrary. Since the physics tells us that many different reference frames or coordinate systems are equally legitimate, only things that hold in all of them can be genuine features of the world, apart from our changeable descriptions of it. And abstract geometric objects are the sorts of mathematical objects that are invariant under such changes in description.

Above, we saw that invariant quantities can indicate the structure of a mathematical space. In physics, too, invariant quantities can indicate underlying structure. They can tell us about the structure of the world, according to a physical theory.

An example: the invariance of the laws under time translations suggests that *time itself* is similarly symmetric—that there is no preferred point in time—just as the space-translation invariance of coordinates indicates that there is no preferred point in Euclidean space. Choosing one temporal origin rather than another is just a conventional choice in description, not an underlying distinction in reality. Another: the invariance of the laws under Lorentz transformations suggests that choosing one Lorentz frame rather than another is just a conventional choice in description, not an underlying distinction in reality. There is no temporal-distance structure in the world, according to the theory of special relativity.

By contrast, in Newtonian physics, there is a temporal-distance structure to the underlying space-time. There is a certain preferred-point structure that is lacking from special relativity. Out of all the pairs of spatially separated space-time points, some are picked out as being *simultaneous*, in any frame. According to Newtonian physics, there *is* a simultaneity structure in the world. The theory is not invariant under transformations failing to preserve this structure.[4]

The inference from invariances in the laws to corresponding symmetries in the world is not conclusive. There could still be a preferred Lorentz frame, say. Yet we tend to infer that there is no more structure to the world than what the fundamental laws indicate there is. Physics adheres to the methodological principle that the symmetries in the laws match the symmetries in the structure of the world. This is a principle informed by Ockham's razor; though it is not just that, other things being equal, it is best to go with the ontologically minimal theory. It is not that, other things being equal, we should go with the fewest entities, but that we should go with the *least structure*. We should not posit structure beyond that which is indicated by the fundamental dynamical laws. As a methodological rule, this is what physics has generally done, and successfully so. This, in turn, suggests that the symmetries in the laws give us genuine *insight* into the world's underlying structure.[5]

---

[4] The laws would not make sense without it. Take Newton's first law: every body continues in a state of rest or uniform motion unless acted on by a net force. This law requires that there be a real distinction between the trajectories of particles at rest or in uniform motion, and the trajectories of accelerating particles. If there were no such distinction in the world (independent of our descriptions of it), then there would be no fact of the matter as to whether a given body is accelerating or in uniform motion. See Maudlin, *Quantum Non-Locality and Relativity*, p. 38.

[5] Note the two kinds of symmetry in play here: symmetries in the laws, given by mappings of solutions to the theory onto solutions, and symmetries in the space-time (or other) structure of the world, given by mappings of the space-time (or other structure) onto itself. The idea is that the former give us reason to infer the latter. See John Earman, *World Enough and Space-Time* (Cambridge: MIT, 1989), p. 46, who considers this a condition of adequacy on dynamical theories.

This gives us a way of figuring out the structure of the world, according to a theory of physics. The procedure is analogous to the one we can use for a mathematical space whose structure is not already familiar. Now we look for invariances in the physical laws:

(1) Take the laws of the theory. Take a system governed by these laws.
(2) Transform one allowable coordinate- or frame-dependent description of the system into another.
(3) See what quantities must remain the same under this transformation, in order for the laws to continue to hold. Repeat for any system governed by the theory and any type of allowed coordinate transformation.
(4) These quantities give the mathematical structure needed to formulate the theory in an invariant, coordinate-independent way.
(5) These quantities then indicate the structure of the world, according to the theory.[6]

In special relativity, the laws say that the speed of light is the same in all inertial reference frames. In order for this to be so, the transformations between different such frames cannot be the classical Newtonian (Galilean) ones. They must instead be the Lorentz transformations. The invariant quantity under the Lorentz transformations—the mappings which take a system's description in one allowable frame to its description in another—is the space-time interval. This space-time interval then picks out a Minkowski geometry for the underlying space-time structure. This is the mathematical structure needed to formulate the theory in an invariant, frame-independent way. Hence this is a part of the space-time structure of the world, according to the theory.

Note that we can also compare different degrees, or *amounts*, of structure. Compare a Euclidean plane with a similar plane that has a preferred spatial direction. The Euclidean plane without a preferred direction has less structure than the one with a preferred spatial direction. Picking out a preferred direction requires additional structure (an orientation).

In building up a mathematical space, some objects will presuppose others, in that some of the mathematical objects cannot be defined without assuming others. Starting from a structureless set of points, we can add on different "levels" of structure. A bare set of points has less structure than a topological space, a set of points together with a topology (specifying the open subsets). A topological space has less structure than a metric space: in order to define a metric, the space

---

[6] I am taking for granted that there is some way in which this mathematical structure represents the world. How this works is of course a large issue, which I leave aside here.

must already have a topology.[7] (Intuitively, a metric gives distances along curves by adding up the lengths of segments between nearby points; and without a topology, there is no sense of the "nearness," or neighborhoods, of points.) And so on.

We can compare the structures of different physical spaces in the same way. The space-time structure of Newtonian mechanics contains a spatial metric, a temporal metric, and structure identifying spatial locations across times; whereas the space-time structure of special relativity is fully specifiable by means of one space-time metric. The latter space-time has *less* structure. It lacks the structure to identify spatial locations at different times, for instance. That would require *additional* structure.

We can now say some general—admittedly imprecise—things about "structure" in physics. Structure has to do with the invariant features of a space, whether an abstract mathematical space or the physical space(time) of the world. Structure comprises the objective, fundamental, intrinsic features, the ones that remain the same regardless of arbitrary or conventional choices in description. And we can learn about structure by looking at the invariant quantities under allowable transformations. I suggest that it is part of the structure of the world, according to a theory of physics, that it contain at least the amount of stuff needed to state the theory in an objective, coordinate-independent way—in a way best suited to representing the nature of the world, apart from our descriptions of it.

Structural features of the world are (somehow: note 6) represented by the mathematical features of the space in which its physics is formulated. But not just any old mathematical features. In relativity, it is the invariant, geometric features, not the frame-dependent, numerical ones, which get at genuine space-time structure. Since not just any mathematical feature indicates underlying structure, we might already be witnessing a glimmer of good things to come. If only certain kinds of mathematical object are appropriate for representing the structure of the world, then this may help us with different mathematical formulations of the physics.

For now, rest content with the following observation. It seems pretty clear why modern physics is so interested in structure. Structure consists in just the kind of things we take to be candidates for objective features of the world, for features of reality. Reality is observer-independent. It does not depend on our arbitrary descriptions or conventions. It is independent of choice of reference frame. Reality has to do with *structure*.

---

[7] For any metric (or pseudo-metric) space, we can define a topology by using the open balls as a sub-basis. The open balls $B_\varepsilon(x) := \{y \in X \mid d(x, y) < \varepsilon\}$ form a basis for the neighborhoods of x, where d is the metric. We say that this is the topology *induced by* the metric. See Chris J. Isham, *Modern Differential Geometry for Physicists* (Singapore: World Scientific Lecture Notes in Physics, 2003, 2nd ed., Volume LXI), p. 33.

II. THE CASE: LAGRANGE V. HAMILTON

We are not out of the woods yet. If structure is supposed to tell us about the objective features of the world according to a theory of physics, then what do we say when there are different structures we can use for the physics?

Let us turn to our concrete case. There are two formulations of classical mechanics that are generally assumed to be equivalent: the Lagrangian and Hamiltonian formulations. Each formulation has its own sets of coordinates and its own equations of motion, given in terms of those coordinates. But the Lagrangian equations of motion can be recovered from the Hamiltonian ones, and vice versa. The two formulations give rise to the same set of physically possible histories for any classical system. The consensus view is that these are notational variants, different mathematical presentations of one and the same theory.

Bear with me for a moment while we run through some details that will be important later. First, some features of the Newtonian formulation of classical mechanics that carry over analogously to these other formulations. In Newtonian mechanics, the total state of a system at a time is given by the position and momentum of each of its particles (in addition to the particles' intrinsic features, like mass and charge).[8] For a system moving around in three-dimensional space, the position and momentum coordinates will each have three components, one along each of the three spatial dimensions.

The *statespace* of a system is a mathematical space in which we represent all its possible fundamental states. Think of this as the fundamental possibility space of a theory. In Newtonian mechanics, a system with n particles will have a statespace of $6n$ dimensions: one dimension for each of the three position and momentum coordinates for each particle in the system.[9] Each point in this high-dimensional space represents a total possible state of the entire system. Each point represents the state of many, many objects—all the particles that make up the system. Keep in mind that a theory's statespace does not in general directly correspond to the physical space of the world. Statespaces tend to look quite different from that.

---

[8] I leave aside the question of whether velocity is properly included in the instantaneous state of a system. I tend to agree with arguments that it is not. (See David Albert, *Time and Chance* (Cambridge: Harvard, 2000), chapter 1; Frank Arntzenius, "Are There Really Instantaneous Velocities?" *The Monist*, LXXXIII, 2 (2000): 187–208.) If that view is right, then just replace the idea of a system's state at a time with what Albert calls a "dynamical condition," the state within any arbitrarily small time interval. What we need is the information required to plug into the dynamical laws to make predictions.

[9] If the system were one particle, we could represent its state by one point in a 6-dimensional mathematical space; this 6-dimensional space represents all the possible

A trajectory through the statespace, a curve through the statespace parameterized by time, represents a possible *history* of the system, a possible sequence of its fundamental states over time. The dynamical laws say which of these curves are possible histories. In Newtonian mechanics, $\mathbf{F} = m\mathbf{a}$, or $\mathbf{F} = \frac{d\mathbf{p}}{dt}$, governs this evolution. There is one such equation for each particle, in each component direction. (It is a vector equation.[10]) We can group these equations into one big version of Newton's law, which says how the point representing the state of the entire system moves through the $6n$-dimensional statespace over time. Given the initial positions and momenta of the particles and the total forces acting on the system (given by a vector function on the statespace), we can (twice) integrate Newton's equation to get a unique solution.[11] A solution is the history of the system, for this initial state and subject to these forces.

Lagrangian mechanics is a bit different. It also specifies a system's state by using two sets of coordinates. It is important, though, that the Lagrangian formulation uses what are called *generalized coordinates*, or generalized positions, labeled q_i, and their first time derivatives, the *generalized velocities*, $\dot{q}_i$ (i ranging from 1 to n). Generalized coordinates are any set of independent parameters that together completely specify the state of a system. Whereas in Newtonian mechanics we use ordinary position and momentum coordinates,[12] here we can use any set

---

fundamental states, all possible position-momentum pairs, for this system. Two particles, and we need 12 numbers to completely specify the state—six numbers for each particle—and so a 12-dimensional statespace. Three particles, 18 numbers, and an 18-dimensional statespace. In general, the statespace of a classical system has dimension $2nr$, where n is the number of particles and r is the number of degrees of freedom, here assumed to be the three spatial dimensions of ordinary space.

[10] Newton's law is $\mathbf{F}(\mathbf{x}, \dot{\mathbf{x}}, t) = m\frac{d^2\mathbf{x}}{dt^2}$, a solution to which is a vector function $\mathbf{x}(t)$, for initial values $\mathbf{x}(t_0)$ and $\dot{\mathbf{x}}(t_0)$ at initial time t_0, which allows us to find also $\dot{\mathbf{x}}(t)$. There is a transformation rule for vectors, which gives their components in one frame from those in another. In order for a vector equation to be frame-independent (in Newtonian mechanics, it is Galilean-invariant), the components of all vectors must transform according to the same equation. In other words, the law must hold component-wise.

[11] I set aside the intriguing cases of indeterminism. See Earman, *A Primer on Determinism* (Boston: Reidel, 1986); John Norton, "The Dome: An Unexpectedly Simple Failure of Determinism," *Philosophy of Science (Proceedings)*, LXXV, 5 (2008): 786–98; David Malament, "Norton's Slippery Slope," *Philosophy of Science (Proceedings)*, LXXV, 5 (2008): 799–816.

[12] The laws still hold in other coordinate systems, but their form will change if we use, say, polar rather than rectangular coordinates. Not so the Lagrangian equations.

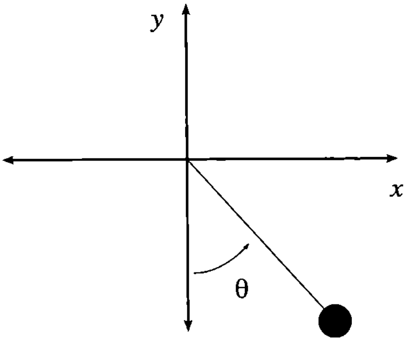


Figure 3: Pendulum with generalized coordinates $q = \theta$ and $\dot{q} = \dot{\theta}$

of generalized coordinates and their corresponding time derivatives; and there are (infinitely) many we could use. In particular, the generalized coordinates need not be ordinary position coordinates. For a pendulum, we can use an angle as the generalized "position"; the generalized "velocity" is the first time derivative of that angle. (See Figure 3.) It is an amazing feature of the theory that we can treat this angle as if it were a rectangular coordinate of the kind we use to solve Newton's equations, and everything goes through just as if we were using ordinary rectangular coordinates.

For a Lagrangian system of *n* particles moving in three-dimensional space, we again need *6n* coordinates to completely specify the state of a system; in this case, three generalized coordinates and three generalized velocities for each particle. Each possible state is represented by one point in a *6n*-dimensional statespace called *configuration space.*

This label is a bit confusing. 'Configuration space' is often used for the space of ordinary positions or configurations, that is, a *3n*-dimensional manifold, where each point corresponds to a possible set of particle positions in ordinary three-dimensional space. Each point in the Lagrangian configuration space, on the other hand, picks out both a generalized position and the corresponding generalized velocity for each particle in the system. This space comprises a set of possible configurations (which, remember, need not be given by ordinary position coordinates), plus a space in which we represent their first time derivatives. This is really the *tangent bundle* of configuration space: the *3n*-dimensional (generalized-coordinate) configuration space together with the *3n*-dimensional tangent space at each point. We need the tangent spaces in order to define the generalized velocities, which are tangent vectors. Unlike ordinary configuration space, the Lagrangian statespace is a *6n*-dimensional manifold, each point of which picks out a generalized coordinate-generalized velocity pair for all the particles in the system.

The Lagrangian equations of motion determine how the point representing the state of the system moves through the statespace over time. Whereas Newtonian mechanics requires the forces acting on a system in order to determine its behavior, here we need a scalar function, called the *Lagrangian, L.* At each point in the statespace, this function assigns a number, (typically[13]) equal to the kinetic minus potential energy of the system. Given the initial state and Lagrangian of a system, the equations determine a unique history.

Hamiltonian mechanics also uses generalized coordinates, but they are different from the Lagrangian ones. The Hamiltonian coordinates are the *generalized positions,* labeled q_i, and the *generalized momenta,* p_i. As generalized coordinates, these need not be ordinary position and momentum coordinates. The generalized momentum need not equal the ordinary momentum of mass times velocity, for instance (and it will not, if the generalized position is not a regular Cartesian coordinate). The Hamiltonian coordinates and equations are called *canonical,* and the statespace is called *phase space.* This is (typically[14]) the *cotangent bundle* of configuration space: the (generalized) configuration space together with the cotangent space at each point. We need the cotangent spaces in order to define the generalized momenta, which are covectors, or one-forms. Once again, this is a *6n*-dimensional manifold; here, each point picks out a generalized position-generalized momentum pair for all the particles in the system. Once again, we need a scalar function on the space in order to predict a system's behavior; here, we need the *Hamiltonian, H,* (typically: note 13) equal to the total energy of the system. For a given initial state, these equations yield a unique history for a system with that Hamiltonian.

---

[13] I ignore time-dependent Lagrangians and Hamiltonians and nonconservative forces, though the theories can accommodate these too. There are many sources for the physics. Some I found particularly useful: Ralph Abraham and Jerrold E. Marsden, *Foundations of Mechanics* (Reading, MA: Benjamin/Cummings, 1980, 2nd ed.); V.I. Arnol'd, *Mathematical Methods of Classical Mechanics* (New York: Springer, 1989, 2nd ed.); Luca Bombelli, chapter II.0 of Abhay Ashtekar, ed., *New Perspectives in Canonical Gravity* (Italy: Bibliopolis, 1988); Herbert Goldstein, Charles Poole, and John Safko, *Classical Mechanics* (Reading, MA: Pearson Education, 2004, 3rd ed.); Victor Guillemin and Shlomo Sternberg, *Symplectic Techniques in Physics* (New York: Cambridge, 1984); Isham, *Modern Differential Geometry for Physicists*; Jorge V. José and Eugene J. Saletan, *Classical Dynamics: A Contemporary Approach* (New York: Cambridge, 1998); Cornelius Lanczos, *The Variational Principles of Mechanics* (New York: Dover, 1970, 4th ed.); Jerrold E. Marsden and Tudor S. Ratiu, *Introduction to Mechanics and Symmetry* (New York: Springer, 1999, 2nd ed.); Dusa McDuff and Dietmar Salamon, *Introduction to Symplectic Topology* (New York: Oxford, 1998, 2nd ed.), and McDuff, "Symplectic Structures—A New Approach to Geometry," *Notices of the American Mathematical Society,* XLV (1998): 952–60; Roger Penrose, *The Road to Reality* (New York: Knopf, 2005), chapter 20; Bernard F. Schutz, *Geometrical Methods of Mathematical Physics* (New York: Cambridge, 1980); Ana Cannas da Silva, *Lectures on Symplectic Geometry* (New York: Springer, 2001); Stephanie Frank Singer, *Symmetry in Mechanics: A Gentle, Modern Introduction* (Boston: Birkhäuser, 2001). See also Gordon Belot, "The Representation of Time and Change in Mechanics," in Jeremy Butterfield and John Earman, eds., *Philosophy of Physics,* Part A (Amsterdam: North-Holland, 2007), pp. 133–227; and Jeremy Butterfield, "On Symplectic Reduction in Classical Mechanics," in Butterfield and Earman, eds., *Philosophy of Physics,* Part A, pp. 1–132.

[14] Hamiltonian phase spaces need not have a global vector bundle structure, though locally they are guaranteed to look like symplectic vector bundles. That their global structure is not similarly constrained is one reason for the conclusion of this paper. See Appendix B for some of the gory details.

The Lagrangian and Hamiltonian formulations are both coordinate-independent versions of classical mechanics. Both are given entirely in terms of generalized coordinates; their equations of motion retain their form regardless of which set of such coordinates we use. (More, they can be formulated without mentioning coordinates at all: Appendix A.) The reason these theories can do this is that the Lagrangian and Hamiltonian functions, which determine the motion of a system, are scalar functions. By contrast, in Newtonian mechanics, the forces determine the motion, and forces are vector quantities. Vectors, we have seen, are coordinate-independent objects. But their components change with the coordinate system, and to solve a problem using Newton's law, we need to know the component forces in the chosen coordinates.[15] The Lagrangian and Hamiltonian functions manage to store this dynamical information in one scalar energy term.

This suggests that the Lagrangian and Hamiltonian formulations are both more objective, description-independent versions of classical mechanics than the Newtonian formulation. In describing a classical system's behavior, these theories allow us to use a wide range of coordinates, even ones that are not ordinary positions and momenta. These theories demonstrate that there is a wider range of variables we can use to describe classical systems than the Newtonian formulation might suggest.

There are differences between the two. Lagrangian and Hamiltonian mechanics use different sets of generalized coordinates to describe systems' states. And not just different coordinates: the generalized coordinates and velocities of the Lagrangian formulation are not related by a canonical transformation—the transformation that takes one allowable set of Hamiltonian coordinates to another—to the canonical position and momentum coordinates.[16] The difference in coordinates means a difference in statespace structure: the structure of the tangent bundle versus the structure of the cotangent bundle. (Though they are both usually fiber bundles, and at that level share an abstract structure.[17])

---

[15] See note 10.

[16] This requires a Legendre transformation, which transforms functions on a vector space (like the tangent bundle) to functions on the dual space (the cotangent bundle).

The equations of motion have notable differences as well. For a classical system with n degrees of freedom (such as the three spatial dimensions we are assuming here), the Lagrangian formulation gives a set of n second-order differential equations, one for each particle in each coordinate direction. The Hamiltonian gives $2n$ first-order equations, two for each particle in each coordinate direction.[18] In each case, though, the equations uniquely determine the motion, given an initial state specified by $2n$ values: for the Lagrangian, the initial n q_i's and n $\dot{q}_i$'s, and for the Hamiltonian, the initial n q_i's and n p_i's.

Despite these differences, it is universally agreed that either formulation suffices for doing classical mechanics. They are each empirically adequate. Each is completely kosher as far as classical mechanics is concerned. Physics books tell us to use whichever set of equations will make a given problem easiest to solve.

It seems we have here a distinction without a difference: a difference in mathematical formulation, not any distinction in reality according to the two theories. The two formulations simply use different variables to describe systems' fundamental states. That is what leads to their differing statespaces and differing equations of motion. We have but one theory, and two different ways of formulating it mathematically, in terms of different, but equally legitimate, coordinates. Just like using polar as opposed to Cartesian coordinates to describe the Euclidean plane.

Mere notational variants indeed.

III. IRRECONCILABLE DIFFERENCES

But not so fast.

Lagrangian and Hamiltonian mechanics may be equivalent for the purpose of doing classical mechanics. Nonetheless, there are important differences between them. There are differences in *structure*.[19]

---

[17] See Appendix B for precisification.

[18] The Lagrangian equations are $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0$, the Hamiltonian $\dot{q}_i = \frac{\partial H}{\partial p_i}$ and $-\dot{p}_i = \frac{\partial H}{\partial q_i}$ (i from 0 to n). See Appendix A for coordinate-free versions. Stating the difference between the two in terms of the order of the equations is a bit misleading, since the Lagrangian equations can be seen as first order equations defined on all of TQ: see José and Saletan, *Classical Dynamics: A Contemporary Approach*, pp. 93–97.

[19] There are also differences that crop up in the extension to other domains of physics. The Lagrangian formulation is generally taken to be more natural for generalizing to field theories and relativistic space-times, the Hamiltonian for statistical mechanics and quantum mechanics (though not quantum field theory). Though standard, none of this is uncontroversial. For one source of disagreement, see Penrose, *The Road to Reality*, chapter 20. I deliberately limit the discussion here to classical mechanics; things might change when we carry these considerations elsewhere.

Look more closely at the structure of their statespaces. The statespace of Lagrangian mechanics—configuration space, the tangent bundle—has metric structure. Lagrangian mechanics assigns a definite, natural geometric structure to configuration space. In particular, the statespace gets naturally associated with a metric structure.[20]

How do we know it has this structure? The usual way: look at what is invariant under allowable coordinate transformations. The invariant quantities of the Lagrangian equations determine the geometric structure of the statespace. They tell us what structure the statespace must have in order to formulate the theory in terms of it.

The Lagrangian equations of motion are invariant under a certain set of (point[21]) transformations. These transformations give the allowable sets of coordinates, the ones that preserve the equations of motion. They indicate which coordinate changes the theory takes to be mere arbitrary changes in description.

There is a quantity that is invariant under these transformations. The invariant quantity is the square of the Riemannian line element. Transformations from one set of Lagrangian coordinates to another—from one set of coordinates giving rise to the Lagrangian equations to another that gives rise to these equations—preserve this local structure, this local curvature, in the form of a quadratic differential form of the $\dot{q}$'s. This allows us to say that the structure of the Lagrangian statespace is that of a Riemannian manifold, with a Riemannian metric defined on it.[22]

The statespace of Hamiltonian mechanics—phase space, the cotangent bundle[23]—is different. Phase space has symplectic structure; it does not have metric structure. Symplectic structure comes with, or determines, a volume element, but not a distance measure.

How do we know it has this structure? In the same way: look at what is invariant under allowable transformations. There is a quantity, a fundamental differential form, that is invariant under the canonical transformations. But it is not the invariant quantity of the Lagrangian transformations. It is the symplectic form, a different sort of geometric object.[24] Transformations from one set of Hamiltonian coordinates to another—from one set of coordinates giving rise to the Hamiltonian equations to another that gives rise to these equations—preserve symplectic structure. So long as a transformation leaves the symplectic structure intact, it can alter any metric structure. Hamiltonian state-spaces that are bent or stretched or "floppy" with respect to one another do not count as different statespaces—not unless the symplectic form differs.[25]

---

[20] Often: not so in the completely general case, though even there, I maintain the basic conclusion about the theories' comparative structures. For ease of exposition, I continue to assume the above statespace structure, relegating technical details and precisifications to the appendices.

[21] In a point transformation, each original point is in one-one correspondence with each transformed point, in such a way that the local structure, in the neighborhoods of the corresponding points, is preserved.

[22] A differential manifold with a fixed positive definite quadratic form on every tangent space is a Riemannian manifold, with quadratic form ds^2 the Riemannian metric. See Peter Szekeres, *A Course in Modern Mathematical Physics: Groups, Hilbert Space and Differential Geometry* (New York: Cambridge, 2004), p. 469, who defines a Lagrangian system as an n-dimensional Riemannian manifold (M, g), together with a function $L: TM \to \mathbb{R}$ (the Lagrangian). Note that there is a more general, coordinate-free version of the theory that does not make explicit use of this metric: see Appendix A. I thank David Malament for this point.

[23] But see Appendix B and note 14 for precisification.

Now here is the kicker. The thing is, as far as we can tell, we need only symplectic structure to do classical mechanics. This structure suffices for the theory; it does not "leave anything out."[26] And there is a clear sense in which a space with a metric structure has *more* structure than one with just a volume element. Metric structure comes with, or determines, or presupposes, a volume structure, but not the other way around. (In the same way that a metric space comes with, or determines, or presupposes, a topology, and not the other way around: note 7.) Intuitively, knowing the distances between the points in a space will give you the volumes of the regions, but the volumes will not determine the distances.[27] Metric structure adds a further level of structure.[28]

---

[24] A symplectic form is a closed, nondegenerate, antisymmetric 2-form. Note that, although the symplectic form determines a volume form, a symplectic form is different from—and stronger than—a generic volume form.

[25] The structure of a symplectic manifold is "floppy" in that there is no local notion of curvature that would distinguish one symplectic manifold from another locally. Two real symplectic manifolds (of the same dimension and signature) are locally identical: they can be mapped onto each other so that their symplectic structures correspond. That symplectic manifolds have no local invariants such as curvature is a consequence of Darboux's theorem, which tells us that every pair of symplectic manifolds is locally isomorphic: within the neighborhood of every point, there are local (canonical) coordinates such that the symplectic form takes the canonical form. See Arnol'd, *Mathematical Methods of Classical Mechanics*, p. 230; Rolf Berndt, *An Introduction to Symplectic Geometry* (Providence: American Mathematical Society, 2001), 2.2; da Silva, *Lectures on Symplectic Geometry*, 8.1. A corresponding result for Riemannian manifolds does not hold unless the manifolds have zero curvature (unless they are flat). Symplectic forms are more flexible than Riemannian metrics, which can be made constant in a local chart iff they are flat. See Abraham and Marsden, *Foundations of Mechanics*, p. 177; Marsden and Ratiu, *Introduction to Mechanics and Symmetry*, p. 148. More on this in Appendix B.

[26] There is an additional topological condition for a symplectic manifold to admit a global Hamiltonian vector field (see Abraham and Marsden, *Foundations of Mechanics*, p. 189); a classical mechanical phase space cannot be compact (momentum can get arbitrarily large): see Singer, *Symmetry in Mechanics*, 3.3. Thus the 2-torus, for example, cannot be a symplectic phase space, even if it happens to be an even-dimensional symplectic manifold. There will also be the Hamiltonian function defined on the phase space. For ease of exposition, I continue to refer to this structure as 'symplectic structure', where this should be understood as including the relevant energy function and any topological conditions.

This suggests that the statespace of Lagrangian mechanics has *more* structure than the theory really needs. We do not need this structure for the dynamics of classical particles; therefore, we should infer that it does not exist, that it does not correspond to fundamental structure. The fact that we can do classical mechanics without the natural metric of the Lagrangian formulation suggests that this extra bit of structure is merely an artifact of the mathematics we need to formulate the theory this way. It suggests that Lagrangian mechanics contains excess, *superfluous structure.*

Note that the Lagrangian formulation cannot get by without this metric structure. The square of the Riemannian line element is what gives the allowable sets of coordinate transformations, the ones that preserve the Lagrangian equations of motion. A volume element would not suffice for this, simply because of how the theory is formulated. The relation between the generalized coordinates and their first time derivatives, the two sets of coordinates that Lagrangian mechanics uses to describe systems' fundamental states, requires this metric.[29] The canonical coordinates of the Hamiltonian formulation do not.

(An example helps to picture the canonical coordinates' relationship. The phase space for a single particle constrained to move along a line, like a bead on a straight wire, is a two-dimensional plane. The symplectic form is the area form, which we can think of as giving the oriented area of the parallelogram spanned by any two vectors. Coordinatize the plane by r and p, interpreted as the particle position and momentum, respectively. For any physically possible motion, along any allowable trajectory, these two coordinates will have a simple kinematic relationship: p will be a constant multiple of the time derivative of r. This relationship is coded up in the area form, the symplectic form for this system's phase space.[30])

---

[27] See Schutz, *Geometrical Methods of Mathematical Physics*, chapter 4. The symplectic form is not even suitable to be a metric, since for any element, the symplectic inner product taken with itself is 0 (from antisymmetry).

[28] I suggest this relationship between the two spaces as a reasonable conjecture; I do not have a proof. See Appendix B for more details.

[29] But see note 20 and Appendix A for precisification.

[30] The canonical symplectic form is then $\omega = dp \wedge dr$; that is, p and r are canonical coordinates. Another example: the phase space of a spherical point pendulum, or a particle constrained to move on a unit sphere. The set of possible positions are represented by points on the surface of a sphere. Since the momentum vector must be tangent to the sphere, the set of all possible position-momentum pairs gives the structure of this phase space as $T^*Q^2 = \{(\mathbf{r}, \mathbf{p}) \in \mathbb{R}^3 \times (\mathbb{R}^3)^* : |\mathbf{r}| = 1 \text{ and } \mathbf{pr} = 0\}$, where Q is the configuration space and T^*Q is the cotangent bundle. See Singer, *Symmetry in Mechanics*, chapter 2.

Of course, this extra bit of structure can be useful. It can help solve problems, or help extend the dynamics to other realms (like field theories). Nonetheless, if our world's fundamental physical theory were the theory of classical particles, we should conclude that the structure of the mathematical space in which to best represent the theory, and of the world according to that theory, has only symplectic structure.

Consider an analogy. Newton thought that he needed a space-time structure identifying spatial locations across times. But this is more structure than the theory really needs. We know that Newton's theory can be formulated in a completely Galilean-invariant way. So we infer that, had Newton's theory been the fundamental theory of the world, its structure would be more accurately represented by Galilean (or neo-Newtonian) space-time, even though Newton believed otherwise.

Similarly here. All the necessary structure for the classical kinematics is contained in the structure of a symplectic manifold—and for the dynamics, with the Hamiltonian function defined on it—even though we can formulate an empirically adequate theory by utilizing more structure.[31] Hence we should infer that symplectic structure is the "real" statespace structure of the theory. We should infer that this structure corresponds to the "real" physical structure of a classical mechanical world. We should infer that Hamiltonian mechanics is *more fundamental* than Lagrangian mechanics.

Another way to motivate the idea. The symplectic form picks out the canonical coordinates by coding up their kinematic relationship; for any allowable motion, the two sets of canonical coordinates must be related in this way. The kinematics is then coded up in the symplectic structure of phase space. The dynamics is given by specifying the Hamiltonian, a function that depends on the particulars of the system in question (the kinds of forces among its particles, any external forces, whether the system is in a potential, and so on). Given the Hamiltonian and initial state of a system, the equations determine a unique history; different Hamiltonians yield different histories. The actual relationship that obtains between the canonical position and momentum coordinates—*which* curve through phase space describes the actual behavior—depends on the initial state and total energy of the system. Finding that relationship amounts to finding a solution, the unique curve describing the behavior of the system. Think of it this way. All the points in phase space represent possible fundamental states, for any system with that number of particles and degrees of freedom; feed in an initial state and Hamiltonian, and the differential equations spit out a unique subset of those points. These points represent the actual development of fundamental states, of position-momentum pairs, for the given system. Or this way: all curves through phase space are functions of q and p; the symplectic form picks out the physically possible such curves; the rest picks out the actual trajectory from among all the physically possible ones.

---

[31] Note that the symplectic form is crucial to determining the motion, since the same Hamiltonian can induce different flows for different symplectic forms. That is, we cannot also drop this amount of structure in an effort to find a more minimal one. At what point we can stop removing structure, while still retaining an adequate theory, is a large remaining question.

The canonical coordinates thus vary in a way that is dictated by the physics. Indeed: the relationships between the generalized positions and the generalized momenta coded up in the symplectic form just *are* the equations of motion themselves. The Lagrangian coordinates are not like this. The generalized coordinates are the only truly independent variables; the generalized velocities are simply their first time derivatives. We treat them as independent for the purpose of solving the Lagrange equations, but they are mathematically constrained, by definition of the time derivative, independent of the physical laws.

Again, this suggests that the Hamiltonian formulation is more fundamental. It suggests that *the* fundamental possibility space for any classical system is a symplectic manifold. Why? Because the relationship between the two sets of canonical coordinates, for any physically possible motion, is contained in the very structure of that space. More boldly: it suggests that a symplectic manifold just *is* the coordinate-independent, geometric representation of classical mechanics.[32] It suggests that classical dynamics is symplectic geometry.

Notice that, on this picture, momentum is a truly fundamental quantity, on a par with position. In particular, momentum is not defined as the product of mass and the time derivative of position.[33] Similarly, it is the points in canonical *position-momentum* space that are fundamental, not the points in ordinary configuration space. (Confession: this is to deny that we should understand the fundamental structure of phase space in the way that it is usually built up, by taking a configuration space and attaching the cotangent space at each point. I am arguing that the *phase space* structure is fundamental.)

---

[32] That is, of the classical kinematics, with the classical dynamics given once we have the Hamiltonian.

[33] This might be to endorse a modernized "impetus theory": see Arntzenius, "Are There Really Instantaneous Velocities?" Since the relationship between the two sets of canonical coordinates depends on a system's Hamiltonian, Arntzenius concludes that, on such a view, this relationship must be regarded as a physical law governing how the canonical position and momentum are causally related; that is, not as a definition of momentum as mass times the temporal derivative of position.

We can now say a bit more about the "structure" of physics. The structure of physics is the (minimal) structure that is assumed among dynamically equivalent—in classical mechanics: canonically transformed—formulations of a theory. Structure comprises those quantities that remain intact when we transform one allowable coordinatization of the laws into another.[34]

In one sense, the orthodox view that the Lagrangian and Hamiltonian formulations are notational variants is correct: the two formulations use different sets of variables to describe the same set of dynamically possible histories for any classical system. But in another, important, sense; they are not *mere* variants: one of them contains *excess structure.* This is a distinction with an all-important difference.

(The defense rests.)

IV. STRUCTURAL REALISM, OR: HOW I LEARNED TO STOP WORRYING AND TRUST THE PHYSICS

When, and why, do we infer the existence of a particular structure, if there are different structures we can use to formulate a theory?

Look to the physics. Take the mathematical formulation of a given theory. Figure out what structure is required by that formulation. This will be given by the dynamical laws and their invariant quantities (and perhaps other geometric or topological constraints). Make sure there is no other formulation getting away with less structure. Infer that this is the fundamental structure of the theory. Go on to infer that this is the fundamental structure of the world, according to the theory.

In classical mechanics, this results in the inference to symplectic structure. This is what remains intact when we change from one allowable coordinate description to another. The theory's alternate formulation tacks on more structure. Infer that the fundamental statespace of classical mechanics is a symplectic manifold. Infer that symplectic structure, the structure that is invariant up to canonical transformations, exists. In a world whose fundamental theory is classical mechanics, infer that it fundamentally contains only this amount of structure.[35]

---

[34] One might wish to say that "structure" has to do with the relations among fundamental objects, rather than their intrinsic properties. I do not wish to put it this way, since I am skeptical of the distinction, for reasons argued for by Maudlin in "Suggestions from Physics for Deep Metaphysics," *The Metaphysics within Physics* (New York: Oxford, 2001), pp. 78–103.

[35] Plus the things mentioned in note 26 and, perhaps, physical space. I do not address here the relation between statespace structure and space-time structure. For now, let us just say: posit all the fundamental structure required for our best physics, where this is not yet to say whether statespace structure is something over and above space-time structure, or whether one of these can be gotten from the other.

The Lagrangian and Hamiltonian formulations thus do not have the thoroughgoing equivalence people readily assume: there is a difference in structure.

I think this gives us grounds for concluding that they are not even genuinely empirically equivalent. One of the lessons of modern geometric formulations of physics is the importance of abstract geometric objects. These invariant mathematical objects, not their coordinate-dependent descriptions, correspond to fundamental, objective features of the world. And we now know that the structure given by these objects can have empirical import. This is most obvious when it comes to space-time structure. In a special relativistic world, there may be no experiment to determine whether there is, or is not, an absolute-velocity structure.[36] Still, we think there is an empirical difference between the two: a difference in the geometric structure of the space-times.

You may not wish to call such a difference "empirical." Ordinarily, we say that there is an empirical difference between two things when there is a possible (idealized, perhaps) experiment to point out that difference. This is not the case here.[37] Some of this is terminological: if you do not like "empirical," substitute "physical," or some other. Yet not entirely terminological. Part of what is to come suggests that there may not be any clear-cut distinction between what counts as genuinely empirical, and not; and in particular, no clear-cut distinction on the basis of something like "in-principle experimental evidence." No difference in possible experimental evidence here, yet even so, we should say there is a difference between the above two space-time structures. I say that we might as well call this an empirical difference: a difference in the way the world is, according to the physics. Call it some other type of difference if you prefer.

Modern geometric formulations of physics suggest that there is more to a theory's empirical content than its set of dynamically possible histories. There is also the statespace in which those histories are traced out. And there is the structure of that space. The equivalence of theories is not just a matter of physically possible histories, but of physically possible histories *through a particular statespace structure.* Hamiltonian and Lagrangian mechanics are not equivalent in terms of that structure. This means that they are not equivalent, period.

Of course, the empirical import of *phase space* structure is not immediately apparent (to say the least). Much easier to see is how points in a mathematical space are supposed to represent the world in relativity: the world has a space-time that is directly isomorphic to the mathematical space representing it. Phase space is much harder to picture. The points themselves represent objects that are not easily conjured up—the dynamically possible states of a system (even the entire world). The structure of a fiber bundle is not directly isomorphic to anything familiar from our experience, such as physical space.

---

[36] But see Albert, "Special Relativity as an Open Question," in Heinz-Peter Breuer and Francesco Petruccione, eds., *Relativistic Quantum Measurement and Decoherence* (Berlin: Springer, 2000), pp. 3–13, for suggestion that there may be such an experiment.

[37] I thank Adam Elga for this point.

Still, I think we should infer that this structure somehow corresponds to the fundamental structure of a world governed by classical mechanics, and for the same reasons we infer a particular space-time structure from the abstract geometric structure of relativity. We should infer that the world contains at least the amount of structure needed to formulate its fundamental dynamics in an invariant way. And physics has generally been successful with the further inference that the world contains no more fundamental structure than that. Just as we take seriously the structure of physical space from the mathematical structure of relativity, we should do the same for the structure of classical mechanics, even at this more abstract, not-so-easily-envisioned level of structure.[38] Just as we take seriously space-time and its geometry, we should do the same for statespace and its geometry.

Note, for all that I have said so far, I am not necessarily a *substantivalist* about phase space, though I confess that I am happy with that view.[39]

---

[38] Other areas of modern physics also suggest that our world's fundamental structure might not be isomorphic to everyday physical space. Consider the fiber bundle structure of modern formulations of classical field theories (see Maudlin, "Suggestions from Physics for Deep Metaphysics"), or the configuration space of quantum mechanics (see Albert, "Elementary Quantum Mechanics," in J.T. Cushing, A. Fine, and S. Goldstein, eds., *Bohmian Mechanics and Quantum Theory: An Appraisal* (Boston: Kluwer, 1996), pp. 277–84).

[39] Why? Two steps. (See David Malament, Review of Hartry Field's *Science without Numbers: A Defense of Nominalism*, this JOURNAL, LXXXIX, 9 (September 1982): 523–34, for argument along both lines.) The first toward an anti-nominalism. Although we may be able to do science without numbers, as Hartry Field (*Science without Numbers: A Defense of Nominalism* (New York: Blackwell, 1980)) argues, it is hard to see how we can do without mathematical objects altogether. For we need the abstract—"abstract" in the sense of "non-coordinate-dependent"—invariant, geometric ones. Our best physics uses these sorts of mathematical objects to represent the fundamental structure of the world, which makes it difficult to see how we could do without them; say, by using the representation theorems Field employs to show that we can dispense with numbers. Classical mechanics is a case in point: the points of phase space themselves represent abstract objects, dynamically possible states. (Similarly for quantum mechanical configuration space.) First conclusion: these mathematical objects are indispensable to our best physics, and this indispensability is the best evidence we can have for these objects' existence. Second, an "anti-abstract/concrete-distinction-ism." Once we start talking about things like "space-time points" at all, the distinction between the abstract and the concrete starts to feel pretty hazy. Second conclusion: once we have phase space structure, there is nothing to stop our going whole-hog "substantivalist" about the

The claim is just this: that phase space is as much a part of the representational content of classical mechanics as the theory's space-time is.

Why take the statespace so seriously? Rather than saying that phase space somehow corresponds to a part of reality, why not regard it as representing the world more indirectly? Perhaps it merely tells us something about the relations among fundamental properties in our world. After all, we do not infer "structures" on the basis of any old relations among physical properties, even if we can model them in a mathematical space with a certain kind of structure. Down that road lies temperature structure, color structure, altitude structure, and on and on and on.[40]

A few reasons, all of them contentious. One: I am skeptical that there are any such clear-cut distinctions between fundamental properties and fundamental relations; worse, between substances and properties.[41] Two: unlike temperature, color, and so on, in this case we are dealing in structures required by our best formulations of the fundamental physics. Three: there is no principled distinction between arguments pressing us toward a seriousness about space-time structure, and arguments pressing us toward a seriousness about statespace structure. Three-and-a-half: if we happen to take space-time seriously by being substantivalists, then we should not object to the mere idea of positing structures among fundamental possibilities, whether unoccupied space-time points or unoccupied statespace points (see note 39). Four: there is no principled distinction between the kinds of considerations singling out a particular space-time structure on the basis of the mathematical structure of relativity, and those singling out a particular statespace structure on the basis of the mathematical structure of classical mechanics. Five and up: see below.

And just as we should infer the existence of the basic entities postulated by our best theories of physics, so too should we infer the existence of the basic *structures* posited by those theories. Call this 'structural realism'.[42] I think modern physics suggests that realism about scientific theories is just structural realism: realism about struc-

---

points as well, whether or not these are rightfully considered "abstract" in a traditional sense. The first step gets us the existence of at least some mathematical objects: those comprising the phase space structure. The second gets us the existence of the phase space points in particular.

[40] I thank Adam Elga for this question.

[41] As argued by Maudlin, "Suggestions from Physics for Deep Metaphysics."

[42] This term was independently used by James Ladyman, "What Is Structural Realism?" *Studies in History and Philosophy of Science*, XXIX (1998): 409–24, to describe a somewhat similar view, though one very differently motivated.

ture. Modern geometric formulations of physics suggest that there is such a thing as *the* fundamental structure of the world, represented by *the* structure of its fundamental physics. There is an objective fact about what structure exists, there is a privileged carving of nature at its joints: along the lines of its fundamental physical structure.

Someone like Reichenbach would disagree.[43] Reichenbach thinks that there is no nonconventional sense to be made of "the" structure of the world. There is the empirical evidence, and there are different ways of describing that evidence: in terms of one space-time geometry, or in terms of another space-time geometry plus a global force field. Neither description is privileged. Neither is a candidate for representing the "true" structure of the world. I do not have much to say against such a view. Insofar as we wish to be scientific realists at all, though, I think we must be structural realists. And structural realism is precisely the denial of the Reichenbachian view. Structural realism denies that different theoretical formulations, utilizing different physical structures, must be on a par—even if they yield the same results for our possible experimental evidence.

The nominalist may also bristle at my talk of the existence of structures. Structures, after all, sound awfully like mathematical objects. Still others will think I have been taking the mathematics too seriously. We must not too blithely take the physics at face value, for some of its tools may be just that—useful tools or devices, but representationally inert.[44]

Yet for all that I have said here, it is open what structure *is*, even though I have indeed argued that this structure is not mere idle wheels. At a minimum, I say that the mathematics of phase space is part of the genuine content of classical mechanics. *This* bit of the mathematics used in our physics cannot be representationally inert: it is inherent to the dynamics. One way to spell this out further would be to be a substantivalist about statespace. Again, I happen to be open to the idea. But we need not go that way. The important thing is that phase space is not just a mathematical convenience. Some part of reality is more or less directly represented by it.

There are two big inferences that I am arguing we should make here. The same considerations leading us to posit a particular space-time structure should lead us also to posit a particular statespace structure; and we should be realists about that structure. You may want to deny either one, or both. I reply that the combination of these inferences yields the simplest, most natural interpretation of our best fundamental physics. You may demur: Surely eschewing additional possibilities and abstract structures is what really yields the better overall theory! On this, we simply disagree.

---

[43] Hans Reichenbach, *The Philosophy of Space and Time* (New York: Dover, 1958).

[44] The phrase is from Brad Skow.

The elusive "structure" of physics is, then, something about the minimal, coordinate-independent stuff needed to formulate an empirically adequate, simple and natural, fundamental physical theory. Structure is about how the basic building blocks—which, if our current theories of physics are on the right track, are abstract, coordinate-independent, geometric objects—fit together to form the world, in such a way that all the parts and their organizational features are put to essential use; in such a way that there is no superfluous structure.

In a classical mechanical world, that structure is symplectic structure.

JILL NORTH
Yale University

APPENDIX A. LOOK, MA: NO COORDINATES!

A.1. Hamiltonian Mechanics. With Abraham and Marsden (*Foundations of Mechanics* 3.3), and others, define a Hamiltonian system as follows.[45] Take a smooth manifold M (the configuration space Q, above) together with a smooth, real-valued function H (the Hamiltonian) on the cotangent bundle of M, T^*M. There is a natural (canonical) symplectic form, ω, on T^*M. H and ω together determine a vector field, X_H (the Hamiltonian vector field), on T^*M. (Nondegeneracy of ω, for finite-dimensional M, guarantees that X_H exists.) This vector field has an associated class of integral curves. Projecting these curves onto M gives the allowed dynamical trajectories. The Hamiltonian equations can be written as $\iota_{X_H} = -dH$, or $\omega(X_H, \cdot) = -dH$, solutions to which are integral curves of X_H. (X_H thus defines a flow on the symplectic manifold, and by Liouville's theorem, Hamiltonian flows preserve the volume form: Hamiltonian flows are canonical transformations.) We say that (M, ω, X_H) is a Hamiltonian system. See Abraham and Marsden, *Foundations of Mechanics*, for proofs that this is equivalent to the form of the equations given in note 18, with canonical coordinates p and q.

A.2. Lagrangian Mechanics. Define a Lagrangian system analogously (see also note 22). A Lagrangian system comprises a smooth manifold M and a smooth, real-valued function L (the Lagrangian) on the tangent bundle of M, TM, where L is regular. Regularity (as defined in Abraham and Marsden, *Foundations of Mechanics*, 3.5) captures the requirement that at each point in the tangent bundle to M, the second partial derivative of L with respect to the fiber (the velocity) is invertible. A curve on M is an allowed dynamical trajectory if, when we lift the curve to the tangent bundle and integrate L along the lifted curve, the result is extremal (relative to nearby curves in M). The Lagrangian equations of motion can then be written in coordinate-free terms (as $\mathbf{L}_{\Delta} \theta_L - dL = 0$; see José and Saletan, *Classical Dynamics: A Contemporary Approach*, 3.4), though we do not need the extra details of that here. See the relevant references mentioned below.

---

[45] I thank David Malament for discussion.

APPENDIX B. CROSS-STRUCTURAL COMPARISONS

The invariant quantity of the Lagrangian equations is the square of the Riemannian line element. This is a quadratic differential form with components of a metric tensor. The kinetic energy term, a quadratic differential form of the generalized velocities, is what gives the definite Riemannian element of this space.

The invariant quantity of the Hamiltonian equations is the symplectic form (with canonical form $\omega = dp \wedge dq$), where the generalized q's and p's are allowed to vary independently. Since H is a function of these two sets of coordinates treated as independent variables, we do not get the invariant quadratic differential form of the Lagrangian transformations. Here, the kinetic energy is a simple linear function of the generalized momenta and positions, considered as independent coordinates.

Add to this Darboux's theorem demonstrating that any two symplectic manifolds are locally isomorphic (note 25), and it seems this should mean that not every Hamiltonian phase space is isomorphic to a Lagrangian statespace, and further, that the kind of structure had by the former is the more general, "looser" structure. Since all symplectic manifolds of the same dimension and signature are locally the same, symplectic forms are more flexible than Riemannian metrics, which can be made constant in a local chart iff they are flat.

The immediate worry is that there is a natural isomorphism between vectors and one-forms, and more generally between tangent bundles and cotangent bundles. If the Lagrangian and Hamiltonian statespaces are both vector fiber bundles—the tangent bundle TM and the cotangent bundle T^*M, respectively, for base manifold M—then we can find a natural dual basis, in which case (since they have the same dimension) they will be isomorphic as vector spaces.[46] Indeed[47] (here including abstract indices and using the usual summation convention), for a finite-dimensional[48] symplectic manifold $(M, \omega_{\alpha\beta})$, nondegeneracy guarantees that $\omega_{\alpha\beta}$ has a unique inverse $\omega^{\alpha\beta}$ (with $\omega^{\alpha\beta}\omega_{\alpha\gamma} = \delta^{\beta}_{\gamma}$), or that the mapping $\omega : TM \rightarrow T^*M$ from tangent vectors to cotangent vectors (with $\omega_{\alpha\beta}v^{\beta} = v_{\alpha}$) is an isomorphism. For a finite-dimensional symplectic vector space (M, ω), nondegeneracy of ω will imply that ω: $TM \rightarrow T^*M$ is an isomorphism. But then how can the two spaces be said to have different amounts of structure?

---

[46] Isham, *Modern Differential Geometry for Physicists*, p. 122.

[47] See II.0 of Ashtekar, ed., *New Perspectives on Canonical Gravity*.

Another worry is the more general version of Lagrangian mechanics (see Appendix A.2), which does not explicitly mention a Riemannian structure on the base manifold. Often, the Lagrangian function has a form requiring a metric—when $L(x,v) = \frac{1}{2}\langle v, v\rangle - V(x)$, where V is a smooth real-valued function on M (the potential)—but it need not have this form. Given this and the above-mentioned isomorphism, how could the two statespaces have differing structures?

I do not have a proof, though I do think it plausible that, even in the most general case, the two formulations will differ in the relevant statespace structure. Two main reasons. First, notice that the isomorphism between TM and T^*M only holds (similarly, the Legendre transform can only be defined) when they are both vector spaces.[49] Therefore, if the Hamiltonian phase space is not at least a vector bundle—where the generalized momenta live in vector space fibers above the generalized-position points of the base manifold—we will not find an equivalent Lagrangian formulation via the usual procedure.[50]

---

[48] That is, for ω strongly nondegenerate. For weakly nondegenerate ω, the induced map between vector fields and one-forms is one-to-one, but in general is not surjective. For the finite-dimensional statespaces here, these will be equivalent, since a linear map between finite-dimensional spaces of the same dimension is one-to-one iff it is onto.

[49] Marsden and Ratiu, *Introduction to Mechanics and Symmetry*, 2.2; da Silva, *Lectures on Symplectic Geometry*, chapter 20.

[50] As in Abraham and Marsden, *Foundations of Mechanics* 3.5–3.6; see also Belot, "The Representation of Time and Change in Classical Mechanics." Here is the gist of that procedure. Take a function L on TQ and solutions to a second-order equation, here the Lagrangian equation. From L we can derive an energy function E on TQ which, when translated to T^*Q by means of the fiber derivative $\mathbf{F}L$: $TQ \rightarrow T^*Q$, yields a suitable Hamiltonian H on T^*Q. (The fiber derivative is the derivative of L in each fiber of TQ, mapping the fiber T_qQ at $q \in Q$ to the fiber T^*_qQ. This is what is often called the Legendre transformation.) The solution curves in T^*Q and TQ will then coincide when projected to Q. An analogous procedure gets from the Hamiltonian to the Lagrangian formulation. In general, if we start with vector bundles over a common base space, then the fiber derivative of a function from L to H, say, will be smooth and fiber-preserving. Note that $\mathbf{F}L$ is not necessarily a vector bundle mapping, but it will be a fiber-preserving smooth mapping: a fiber-preserving isomorphism. This transformation requires that the Lagrangian and Hamiltonian functions be hyperregular. (Abraham and Marsden, *Foundations of Mechanics*, pp. 218–23; Marsden and Ratiu, *Introduction to Mechanics and Symmetry*, 7.4. A Lagrangian is hyperregular iff $\mathbf{F}L$: $TQ \rightarrow T^*Q$ is a diffeomorphism; a Hamiltonian is hyperregular iff $\mathbf{F}L : T^*Q \rightarrow TQ$ is a diffeomorphism. In other words, the fiber derivative will be locally invertible iff L (or H) is regular. Hyperregularity is

And, in general, a Hamiltonian phase space need not have the structure of a vector bundle. Though phase space is often the cotangent bundle of a configuration space,[51] it need not be possible, in a given Hamiltonian phase space, to separate the position and momentum coordinates in this way: they can be freely mixed by symplectic transformations.[52] Darboux's theorem, after all, says that every symplectic manifold is *locally* like a symplectic vector space, in suitable local coordinates. That is, locally, any symplectic phase space will look like a vector bundle. Globally, though, it can have a different structure, for the phase space might not be naturally separable in this way.

This is unlike the Lagrangian statespace, which must be a vector bundle. We build this space up by starting with the configuration space Q as the base manifold, and attaching the tangent spaces, on which we define the $\dot{q}$'s, at each point. This, after all, is how the generalized velocities are defined. The resultant structure does yield a natural symplectic form, which may lead you to ask: Why not prefer the Lagrangian formulation, where we seem to get the symplectic structure for free? Answer: the most general structure for a classical system's phase space is a kind of structure that need not be easily or naturally splitable into a configuration space plus associated tangent spaces in this way.[53] The Lagrangian statespace structure is therefore a more restrictive, less general structure.

---

thus needed for the global result.) Hyperregularity is needed for the relevant map from Lagrangians to Hamiltonians to be a bijection. Hyperregularity amounts to the requirement that the underlying statespaces be isomorphic in this way; and without it, we cannot use this procedure to generate a Hamiltonian from a Lagrangian or vice versa.

[51] And it can be shown that cotangent bundles are equipped with a canonical symplectic structure: Abraham and Marsden, *Foundations of Mechanics*, 3.2.

[52] Any $2n$-dimensional vector space Γ can be regarded as a cotangent bundle by assigning two complementary n-dimensional subspaces, Q and P, with $Q \cap P = \{0\}$, $Q \oplus P = \Gamma$. See chapter II.0 in Ashtekar, ed., *New Perspectives on Canonical Gravity*. When Γ is also a *symplectic* vector space, the symplectic form ω picks out a preferred class of pairs of complementary subspaces, namely the ones with $\omega_{\alpha\beta}v_1^\alpha v_2^\beta = \omega_{\alpha\beta}w_1^\alpha w_2^\beta = 0$, for all $v_1^\alpha, v_2^\beta \in Q$ and $w_1^\alpha, w_2^\beta \in P$. Then we can choose convenient bases $\{v_\mu^\alpha\}$ and $\{w_\mu^\alpha\}$ in Q and P, respectively, such that $\omega_{\alpha\beta}v_\mu^\alpha w_\nu^\beta = \delta_{\mu\nu}$. With any such choice of basis, the symplectic form will have the standard (canonical) form, and we have effectively identified configuration and momentum variables in Γ: we can write $v_\mu^\alpha = (\partial/\partial q^\mu)^\alpha$ and $w_\mu^\alpha = (\partial/\partial p_\mu)^\alpha$. This, however, is a special case. Here, the state of a system can be specified by means of its configuration variables q^i and momentum variables p_i, where there is this natural distinction between the two: the q^i's are coordinates on a configuration space Q, and for each q^i, the p_i's belong to the linear space of cotangent vectors at q^i (in a coordinate system). Here the phase space has a cotangent bundle structure, T^*Q. In a general Hamiltonian phase space, however, this natural splitting of Γ is not available: the phase space need not always be a cotangent bundle. An example of the more general case (from Ashtekar, ed., *New Perspectives on Quantum Gravity*, p. 22) is given by the choice of $\Gamma = S^2$, $\omega_{\alpha\beta} = \varepsilon_{\alpha\beta}$, a volume 2-form.

[53] The induced symplectic form of the Lagrangian formulation requires a vector bundle structure. Peter Forrest suggested to me this characterization of what is going

So there is a natural isomorphism between the two statespace structures. But of course there is, if and when they are both vector fiber bundles. In that case, there will be an isomorphism that preserves the vector bundle structure: a vector bundle isomorphism. But this is not sufficient to show that the two spaces possess the same amount of structure. As a general rule, not just any isomorphism will be the relevant structure-preserving map. (A "cardinality-isomorphism" would often leave out much to be desired.) And this is one of the cases following that rule.

Once again, symplectic structure comes out as more fundamental, permitting just those sets of coordinates that allow for the equations of motion of classical particles to hold, regardless of whether those coordinates are perspicuously related to ordinary position and momentum coordinates. Symplectic structure is a simpler, more natural, more general underlying structure—but a structure that suffices for the theory all the same.[54]

Second, look at the underlying groups.[55] Start with the linear case of a $2n$-dimensional vector space, call it V. If V is equipped with a positive-definite symmetric bilinear form (the linear analog of a Riemannian metric), then the group of linear transformations preserving this structure is $O(V)$, a group with dimension $2n^2 - n$. If V is equipped with a symplectic form, on the other hand, then the group of linear transformations preserving the structure is $Sp(V)$, a group with dimension $2n^2 + n$. The former is a stronger structure, in that it admits a smaller group of symmetries. The difference magnifies for manifolds. The largest symmetry group of a Riemmanian metric on a $2n$-dimensional manifold is $2n^2 - n$, whereas every symplectic manifold admits an infinite-dimensional group of symmetries.

To put it another way, the Lagrangian equations of motion are invariant under a set of point transformations; the Hamiltonian, under

---

on. Despite the natural isomorphism between the two statespaces, there remains the following order of explanation: the symplectic form on the phase space is more natural, since the induced symplectic form on the Lagrangian configuration space is only definable by way of the Hamiltonian.

[54] Another consideration mentioned to me by Frank Arntzenius, which warrants further investigation. The Lagrangian formulation seems to presuppose a metric on time. (Recall its formulation via least action principles, where we integrate the action term with respect to the time along a path.) The Hamiltonian, it seems, does not. (Consider the Poisson bracket formulation, independent of the time coordinate and equivalent to the canonical equations: Abraham and Marsden, *Foundations of Mechanics*, p. 198.)

[55] See Abraham and Marsden, *Foundations of Mechanics*, chapter 9. For discussion, I thank Gordon Belot (who disagrees with my conclusions).

the canonical transformations. Whereas all point transformations are canonical transformations—point transformations form a subgroup of the set of all canonical transformations[56]—point transformations are orthogonal transformations, and thus only a special type of canonical transformation. Canonical transformations are more general, precisely because the generalized momenta are treated as truly independent of the generalized positions. As Herbert Goldstein, author of the classic textbook on mechanics, puts it:[57]

> The advantages of the Hamiltonian formulation lie not in its use as a calculational tool, but rather in the deeper insight it affords into the formal structure of mechanics. The equal status accorded to coordinates and momenta as independent variables encourages a greater freedom in selecting the physical quantities to be designated as "coordinates" and "momenta." As a result we are led to newer, more abstract ways of presenting the physical content of mechanics.

---

[56] Abraham and Marsden, p. 181.

[57] Goldstein et al., *Classical Mechanics*, p. 369.

PROBABILITY IN DETERMINISTIC PHYSICS*

In a deterministic theory, one can deduce the state of the universe at any one time from the dynamical laws that govern its evolution and its state at any other time. In particular, one can deduce from the conditions that obtained at some early boundary condition for the universe its state at all subsequent times. It is a piece of well-received philosophical common sense that there is no room in a deterministic theory for objective probabilities. Arguments will be given below that this piece of philosophical common sense is mistaken, and that a probability measure over the space of physically possible trajectories is an indispensable component of any theory—deterministic or otherwise—that can be used as a basis for prediction or receive confirmation from the evidence.

Phase space is the central representational space in physics. It is the space within which all of the states pertaining to a system are represented, and it encodes all of the general information that we have about those states. To take an obvious example (less obvious examples will be given below) the fact that a pair of quantities are represented by different dimensions in phase space encodes the fact that they represent separate degrees of freedom. Physically possible histories are represented by trajectories through phase space and in a deterministic theory, when dealing with a closed system, there is (no more than) one trajectory through each point. Since the universe as a whole is the only truly closed system, this means there is one trajectory through each point in the phase space of the universe. Probabilities may be invoked in applications of theory, when the state of a system or the values of external variables are not known with full precision, but they are eliminable given complete knowledge. And they do not appear in the statement of the laws. The laws serve simply to identify the physically allowable trajectories. About Newtonian dynamics, Laplace famously said

> given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements

---

*Many thanks to Guido Bacciagaluppi, Huw Price, Elliott Sober, and especially Brien Harvey for comments. And I am grateful to audiences at Brown University, Sydney University, and the University of Western Ontario where earlier drafts of this paper were presented.

0022-362X/09/0602/89–108

> of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, and the past, would be present to its eyes.[1]

It is a short step from this to the idea that probabilities invoked in classical contexts are *subjective* rather than *objective*. If they represent anything, it is facts about our epistemic states, not facts about the world.[2] Karl Popper, and David Lewis, and other influential figures have endorsed this view, categorically rejecting any role for objective probabilities in deterministic contexts. In Popper's words:

> Today I can see why so many determinists, and even ex-determinists who believe in the deterministic character of classical physics, seriously believe in a subjectivist interpretation of probability: it is in a way the only reasonable possibility which they can accept: for objective physical probabilities are incompatible with determinism.[3]

And it is not just the philosophers; physics textbooks standardly assert that it is only with the advent of quantum mechanics where the laws take an explicitly probabilistic form that probability has to be taken seriously as part of the objective content of the theory.[4] Barry Loewer,[5] David Albert,[6] and others have been laboring to obtain recognition for the objective character of the probabilities given in the microcanonical probability distribution in Statistical Mechanics, and Elliott Sober has argued for the recognition of objective macroprobabilities.[7] But these have not gained a wide recognition for the role of probability quite generally. This paper will argue from general considerations that any theory capable of receiving confirmation from the evidence or generating expectation in a realistic setting incorporates a measure over state space as an indispensable component of its objective content. Contrapositively, a theory devoid of such a measure is practically devoid of empirical content and practically incapable of receiving evidential support.

---

[1] Pierre Laplace, *A Philosophical Essay on Probabilities* (1814), F.W. Truscott and F.L. Emory, trans. (New York: Kessinger, 2007), p. 4.

[2] Barry Loewer, expressing but not endorsing this reasoning, has written:

"If the laws are deterministic then the initial conditions of the universe together with the laws entail all facts—at least all facts expressible in the vocabulary of the theory. But if that is so there is no *further* fact for a probability statement to be about"—"Determinism and Chance," in *Studies in the History of Modern Physics*, XXXII, 4 (2001): 609–20, on p. 609.

[3] Karl Popper, *Quantum Theory and the Schism in Physics* (New York: Rowman and Littlefield, 1985).

[4] Quantum probabilities will be discussed in the sequel to this article. The laws of quantum mechanics specify not what *will* happen, given some initial segment of history, but what *might* happen with such and such probability.

[5] Loewer, "Determinism and Chance," and "David Lewis' Humean Theory of Objective Chance," *Philosophy of Science* LXXI, 5 (2004): 1115–25.

[6] Albert, *Time and Chance* (Cambridge: Harvard, 2000).

[7] Sober, "Evolutionary Theory and the Reality of Macroprobabilities," in Ellery Eells and James Fetzer, eds., *Probability in Science* (Chicago: Open Court, forthcoming).

There is a very lively debate about the nature of probability (whether they are objective or subjective, whether they supervene on nonprobabilistic or nonmodal facts, and so on). A two-step approach will be favored here, separating what one says while working in the context of a physical theory where notions like law and probability and space and time are treated as primitive, from and what one says *about* laws or probability or space or time in a meta-scientific voice, where attempts at reduction or analysis can be offered. Aside from a few brief comments to address potential concerns of Bayesians, the discussion here will stop at the level of physical theory. The suggestion will be simply that we should recognize a probabilistic postulate as one of the standard components of a theoretical package, alongside the laws and space of states.

WHAT DETERMINISM ENTAILS

It was remarked above that there is only one physically possible trajectory through the phase space of a universe governed by deterministic laws. This means that if the state of the universe is known with perfect precision at any time, its state at any other can be predicted with certainty. In physics, however, when one is interested in the universe as a whole, one never knows its state with perfect precision, so it is represented not with a point, but with a finite volume of phase space. When one is not interested in the universe as a whole, but in an open subsystem of the universe, one is still dealing implicitly with finite volumes of universal phase space. For the state of an open subsystem of the universe is a partial specification of the state of the universe as a whole. Anything larger than a point-sized volume in the phase space of such a system corresponds to a finite volume in the universal phase space (extended in all dimensions external to it). When one is dealing with finite volumes of phase space, even in a universe governed by deterministic laws, one is no longer in a situation in which there is a single trajectory leading from one's starting point. There are multiple trajectories—typically, an infinite number—coming out of any finite volume of phase space. And dynamical laws that constrain only possibility—even deterministic laws that yield a *single* possibility for every point in phase space—are impotent to discriminate between possibilities. They will tell us which trajectories are possible, but will be silent about how to divide opinion among them.

In practice, when one is, for example, carrying out an experiment on some localized system in the lab, one can ignore the possibility that a comet will come at light speed and destroy the apparatus, that the air in the room will anti-thermodynamically collect in one corner, or that the pointer on our measuring device will be disturbed by a quantum tunneling event, even though there are models of the physical laws in which each of these occurs, because these possibilities have negligible or low probability. As possibility goes, they are on a par with those we take seriously (for example, that the result we observed tells us something about the system we are measuring). The epistemic symmetry is broken only by relative probability. It does not help to say there that many open subsystems are *approximately* closed, for probabilities are needed to make out *degrees* of closure. To say that a system is approximately closed is to say that the *probability* of outside influence is minimized.

If probabilities are needed to guide expectation under limitations in knowledge of the state of the universe, they are needed also to relate theory to evidence where it is a question not of *logical* compatibility, but *degree of support*. Logical compatibility with the data is a very weak constraint, certainly not strong enough to underwrite the theoretical choices that scientists actually make. In practice, theoretical choices between alternatives—from mundane, everyday inferences to choices between competing scientific theories—always involve assessments of likelihood. The Law of Likelihood says that given a choice between a pair of theories H_1 and H_2 compatible with a body E of evidence, one should choose the one that assigns the E a higher probability. Objective assessments of likelihood make use of a measure that the theory itself provides. In assigning a likelihood to H_1, one asks how special or contrived the initial conditions have to be by H_1's own lights to generate the regularities present in the phenomena. In assigning a likelihood to H_2, one asks how special or contrived or improbable the initial conditions have to be by H_2's lights to generate the regularities present in the phenomena. How much of what is observed in fact is inevitable, and how much is the result of coincidence or accident? If we observe a robust correlation between a pair of variables where a theory recognizes no connection, that lowers evidential support for the theory not because the theory rules it out as a logical matter, but because the theory assigns it a low probability. It is unlikely, but not impossible, that a pair of independent variables exhibit correlations. In the case of an indeterministic theory, we ask how likely are those regularities, given the chances described by the laws. If a theory assigns a high probability to one result, but whenever we carry out the experiment we always get the opposite, this disconfirms the theory not, again, because of a logical incompatibility, but because the theory assigns that result a low probability. The likelihood of the theory is a measure of how improbable the evidence is, by its own lights, and in making relative assessments of likelihood, we always employ a probability measure.[8]

Probability here is playing an indispensable mediating role between theory and evidence. Any theory that receives confirmation that favors it over empirically adequate alternatives on the basis of likelihood reasoning, or generates expectation that discriminates possibilities, has to tell us not just what is physically *possible* relative to a body of evidence, but the degree to which the various possibilities are to be *expected*. The kinds of regularities that are actually used, even in the most highly regimented experimental contexts (to say nothing of how things are out in the wild where expectations are formed about radically open systems on the basis of only the roughest information) are regularities about the probability with which a system that starts out in some finite, more or less well-defined, volume of phase space and is potentially subject to an indefinite number of potential extruding factors, ends up in another. Even if the intrinsic state of a local system is known with precision arbitrarily close to perfection, that state fills a finite volume of the phase space of the universe as a whole. In practice to derive expectations for a system *S* what is done is that one writes down everything that is known and thought to be relevant about *S*, about its environment, about relevant influences on it.[9] One comes up with a physical description as precise as possible in relevant respects.[10] That will carve out the volume of phase space of interest. He then takes his expectations for *S* from the behavior of systems whose states fall in that subspace, that is, systems that share everything that is known about *S*.[11] So, for example, if one wants to know how a particle in a certain state will evolve under a specified set of conditions, one identifies the subspace of the universal phase space that corresponds to particles of that kind, identifies the volume of that subspace that corresponds to the relevant state, and applies the dynamical laws to determine how states that fall in that volume evolve under the relevant conditions. In doing so, one is effectively assigning to the system of interest the profile of the typical ensemble compatible with everything he knows about it. This fixes the values of known variables and randomizes over unknown variables. The probability of *b* on *S* calculated in this way is the probability of a random pick from a typical ensemble of *S*-type systems.

---

[8] Employing subjective probabilities in this capacity makes relative likelihood assessments a subjective matter and renders it all but void as an objective basis for choice between theories. If the measure is included in the scope of the theory as part of its content, the Law of Likelihood can provide a theory-independent measure of degree of fit with the evidence. See my "Likelihood Reasoning and Theory Choice" (manuscript).

[9] It is not clear that the notion of relevance can be made out without invoking probability, for *A* is a relevant factor if there is a non-negligible probability of influence from *A*. But put this aside.

[10] This is an application of the Principle of Total Evidence, usually credited to Rudolph Carnap, *Meaning and Necessity* (New York: Clarke, 1947).

[11] To know someone is a typical member of a flatly distributed population leaves one without information that favors one location over another, to know someone is a typical member of an unevenly distributed population is to know something that favors some locations, but not enough to locate them in a subpopulation. A typical American, for example, is more likely to live in New York than Tucson, but a typical molecule of water in the ocean has roughly equal probability in being located in any two volumes of equal size.

Even in a deterministic context, in which there is only a single trajectory through every point in phase space, unless one knows the state of the universe as a whole with perfect precision, probabilities are unavoidable. In practice one is warranted in ignoring most of the universe when considering the state of a local system because she has a probability distribution, at least implicit, that assigns a low probability to distorting influences from other parts of the world, that is, a probability distribution that defines what is meant by "random" and licenses the treatment of variables not represented in the model as random. That same probability distribution lets one interpret what she does observe. If one gets the same result every time an experiment is carried, although it is *possible* that this is pure coincidence, or that there is a string of experimental errors (the lab assistant misread the result, there is a freak event inside the measuring apparatus, and so on ...), these possibilities have low probability. The more reliably a result is observed, and the more robust it is with respect to variation in the surrounding circumstances, the more significant it becomes because, and only because, the probability of coincidence or accident is lowered if the result persists over time and across variation. The possibility of coincidence or accident remains, it is only their probability that is diminished. Again, the probabilities are playing a crucial mediating role between evidence and theory.

The probabilities here are independent of the dynamical laws. If one is looking at a finite volume v of state space, there will almost always be multiple trajectories coming out of v and leading in different directions. Some will go up and some will go down, and to form expectations, one needs to know *how many* of a typical finite set of systems whose states begin in v go this way and *how many* go that. If one wants to know the probability that a system whose initial state falls in v is in v^* after an interval *I*, that is to say, she calculates the probability that a random pick from a typical ensemble of systems whose states begin in v, ends in v^*. Dynamical laws that specify only what the physically *possible* trajectories are will not yield these probabilities. And so where there is more than one possibility, dynamical laws will not specify how to divide opinion among them.

Since, in the deterministic case, at least, the need for probabilities disappears when one possesses perfectly precise knowledge, it is easy to slip into thinking that the probabilities are subjective. If probabilities are only *needed* in forming expectation where there is ignorance, the probabilities must express facts about the epistemic states of belief-forming agents, rather facts about the systems they are forming opinions about. It is true that the probabilities reflect facts about the epistemic states of belief-forming agents, but there is a subtle equivocation in passing from the thought that probabilities provide a measure of degree of ignorance to the idea that they express nothing more than degree of belief. Credence is not the same as degree of ignorance. To characterize one's states as states that reflect a lack of information, one needs to invoke a notion of probability that is distinct from credence. A flat distribution over a partition of possibilities expresses ignorance only if the cells of that partition are equiprobable. Suppose I am attending a hosted dinner and I am uncertain about what will be served. If I assign equal credence to anything in the cookbook, that expresses ignorance about what my hosts will cook. But I could equally partition the possibilities for dinner into Sea Bass on quinoa pilaf on the one hand and anything other than Sea Bass on quinoa pilaf on the other, and an equal assignment of credence to the cells in this partition expresses quite specific information about the menu. Ignorance is indifference *between equally probable alternatives*, and the credences one adopts about S when one does not have specific information about S will inevitably invoke an objective, statistical notion of probability and express beliefs about which alternatives are equally likely to show up in a random sample. Credence that favors A over B or B over A, likewise, expresses the possession of information only if the expectation of A and B would be otherwise equal. A probabilistic postulate provides a base-line definition of statistical randomness that allows one to characterize credences as states of knowledge or ignorance.

These probabilities form part of the empirical content of the theory. It was remarked above that in forming expectations for a system S, one takes expectations from the behavior of typical systems compatible with everything known about S.[12] Facts about the behavior of typical ensembles of systems whose states span a finite volume of phase space are perfectly objective. They are probabilistically related to frequencies, and confirmable or disconfirmable by collecting statistics from ensembles. They are as objective, and as empirical, as facts about the phases of the moon, or the migration patterns of butterflies. "But surely," one might say, "deterministic dynamical laws dictate how any ensemble would evolve. If one knows the initial states of the systems of which it is composed, one applies the dynamical laws to determine their final states, and one is done." That is correct, the problem is not that one does not generally know how an ensemble of some specified composition will evolve, the problem is that one does not know *the composition of the typical ensemble.* This is a crucial difference. And notice that the same will be true in the indeterministic case. The dynamical laws there will generate a probability distribution over properties of interest for a system whose pure state is known, but they will not say how to mix probabilities for different states where there is uncertainty about which of them a system is in.

---

[12] This is the source of the link between single-case and in general probabilities, that is, between expectations for behavior of individual systems and frequencies of behaviors in ensembles.

This is perhaps the deepest lesson of statistical mechanics. One does not get a probability distribution *out* of dynamical laws for free; one has to start with a distribution and evolve it forward. Most often an initial distribution is obtained by giving a flat distribution over the volume of phase space compatible with what we know about a system, tacitly applying a principle of indifference to states represented by equal volumes thereof. But we can reparameterize the space without affecting the topological relations among states (and hence without affecting Boolean relations among their extensions) or dynamical laws. And applying the same procedure under different parameterizations can give wildly inconsistent results. Talk about a flat distribution makes sense only once the metric of phase space has been fixed, and the practice of giving a flat distribution relative to a standard metric to reflect a lack of knowledge just expresses the assumption that states that occupy equal volumes under the standard metric are equally *intrinsically* probable, that is, that a random pick from a population that spans any two subspaces of equal volume is equally likely to yield a system in one as in the other.

The role that the measure over phase space was playing in generating probabilities in statistical mechanics was exposed when it was demonstrated that one can hold fixed the laws governing the evolution of individual microstates and derive quite different probabilities for evolution on the macrolevel (where macrolevel descriptions are just less than perfectly precise descriptions of microbehavior) by inflating some macrostates and deflating others. Statistical mechanics taught us to see the metric as an independent theoretical variable—one of the things that can be played with in bringing theory into line with the evidence. What one is playing with when one entertains different hypotheses about the relative volumes different states take up in phase space is hypotheses about the composition of typical ensembles. The naïve expectation would be that the typical ensemble is a roughly even composition of macrostates.

The hypothesis of microuniformity, by contrast, gives us a composition so heavily weighted toward equilibrium states that almost all systems in almost all possible universes are in equilibrium.[13] Under application of the dynamical laws, the hypothesis of microuniformity yields a migration towards higher entropy states from lower entropy states as byproduct of the difference in phase space volume of states of high and low entropy. Overall, there is almost no migration since virtually all systems are in states of high entropy. But if our universe is (as the Past Hypothesis invites us to suppose) one of those very rare universes in which there are a great many systems in low entropy states, migration will be observed, and the migration observed will virtually all be in the same direction.

So to reconnect with the themes above, according statistical mechanics the expectation that one should see as many entropy decreasing processes as entropy increasing ones depends on a faulty assumption: namely, that all macrostates have equal statistical probability in the sense that a random pick from a typical ensemble is equally likely to yield a system in any chosen macrostate. There is dispute about the status and interpretation of statistical mechanical probabilities. But one thing that can be uncontroversially learned from that example is the purely formal fact that one can get radically different expectations by adopting different metrical assumptions, even while holding the dynamical laws fixed. Dynamical laws alone, even deterministic laws that dictate the evolution of a closed system given precise information about its state, do not constrain the division of opinion when initial information is incomplete or less than perfectly precise, or when dealing with an open system with limited knowledge of extruding factors. Talk of dividing opinion evenly among the possibilities—which here takes the form of spreading it uniformly over the part of phase space whose boundaries reflect our information about the system—is well defined only once the metric has been fixed, and only because the metric is functioning tacitly as a measure of the relative probability that a random pick from a population whose states fall within two specified subspaces will yield a system in either.

Although the lesson was one that historically, emerged from statistical mechanics, it has nothing specifically to do with dynamics, and the point can be illustrated with simpler examples. The need for an independent metrical assumption has to do with mixing, and can be made in nondynamical terms. Whether dealing with points in phase space, cells in a partition, or categories in a taxonomy, dynamical laws that give profiles for individual cells need to be supplemented with a comparative measure of the sizes of the populations that occupy cells, something that specifies how systems distribute themselves among cells and can be used to yield the probability that a random pick (from the population as a whole, or a subpopulation drawn from a specified collection of cells) will yield a member of the cell in question. One gets different expectations depending on which is regarded as partitioning them into equally probable classes.[14] It is much clearer in finite examples where one can simply count up the inhabitants. In these cases, the size of a cell is given by the number of its inhabitants and probability corresponds to proportion. Let us take the set of people in the world and partition them by *country of birth.* Suppose that nationality is deterministic with respect to language and car make, so that the country in which a person is born determines the language she speaks and the kind of car she drives. Everybody born in the United States speaks American and drives a Ford, everybody born in France speaks French and drives a Peugeot, everyone born in Samoa speaks Samoan and walks to work, and so on. In forming expectations about a fellow of whose country of birth one has no specific knowledge,[15] if one naively divided opinion evenly among the possibilities for country of birth, she would expect him to be equally likely to speak Chinese as Samoan, equally likely to drive a Ford as a Peugeot, and so on. She would be mistaken, and mistaken for the simple reason that these probabilities would not reflect statistics of the properties of interest in the world population. Taking the number of people in each cell as a measure of its size, let us say that a partition is *even* with respect to a given population just in case it divides the population into cells of equal size. The problem with the partition by country is that it is not even with respect to world population. A typical randomly chosen ensemble of world individuals is not composed of nationalities in equal proportion, and probabilities guided by assigning equal probabilities to nationalities will not reflect statistics in the global population. Consider by contrast the partition of the same population by date of birth. This is not perfectly even with respect to population, but it is more nearly so.[16] On any given day, the number of people born in China will radically outweigh the number born in Samoa, the number of people born in the United States will radically outweigh the number born in France. The expectations formed by spreading opinion evenly over the cells of this partition will be quite different than those formed by spreading it evenly over the cells of the partition by country, and will be more in line with the actual frequencies.

---

[13] At all times, the expectation is high entropy. If one is in one of those rare universes in which there are many systems in low entropy states, they are almost all evolving to states of higher entropy. Higher entropy states are intrinsically so much more likely that if one is in a high entropy state one is almost certain to stay there, and if one finds oneself in a low entropy state, one is almost certain to evolve into a higher one. The composition of any random ensemble of systems is heavily weighted towards high entropy states.

---

[14] It does not matter whether we are talking about dynamical laws or a function from cell number to properties of interest. It does not matter whether we are talking about the deterministic or indeterministic case. And it does not matter whether we are talking about finite or infinite populations. In the finite case, size means number; in the infinite case, it means measure, but comparisons of size are no less dispensable in the second case than in the first.

[15] That will, in its turn, generate profiles for people of whom we have information represented by a probability distribution over some finite region. The typical global individual provides a kind of standard or reference point, departures from typicality represent knowledge.

So, to repeat, when forming expectations for a system *S* about which one has imprecise knowledge, one writes down everything known about *S* and takes expectations for *S* from typical ensembles of systems that share those properties. A complete profile for members of each cell does not give a complete profile for populations drawn from multiple cells until the proportion that individual cells contribute to the combined population is specified. One can derive different expectations for the global population, or any subpopulation that spans cells, by fiddling with how many elements each cell contributes to the larger population. Talk of metrics for phase space is just a generalization of this. The probability with which an arbitrary pick from the population yields a system in a particular state is proportional to the volume it occupies in phase space. One can use whatever numbers one likes to parameterize phase space, but nonmetric preserving reparameterizations will alter expectations under this procedure for forming them. What is the right parameterization? That will depend on how systems actually distribute themselves over phase space. One has to make some metrical assumption before expectations can be formed. Expectations derived from a flat distribution over a space whose metric does not reflect the way that the systems actually distribute themselves leaves one with probabilities distorted in favor of sparsely populated cells.

Another example of failing to appreciate the need to take account of the relative sizes of the cells of a partition in forming expectations, interesting because it actually reverses conditional expectations for a combined population that hold individually for each of the components. The University of California, Berkeley, was very famously sued for bias against women applying to grad school. The admission figures showed that men applying were more likely than women to be admitted, and the difference was significant. But examining individual departments found most departments actually showed a slight bias against *men*. There is no paradox here. The explanation was simply that women tended to apply to departments with low rates of admission, while men tended to apply to departments with high rates of admission. The population as a whole (people who apply to grad school) was composed of a set of subpopulations (people who apply to the English department, people who apply to the Physics department, and so on). In each subpopulation, the rate of acceptance for men was lower than that for women, but in the combined population, the rate of acceptance for women was significantly lower than that for men because the composition of the ensemble of women applicants was heavily weighted towards departments with low rates of acceptance, while the composition of the ensemble of male applicants was weighted towards departments with higher rates of acceptance, with the result that men had a higher overall probability of being accepted. It is as though they were in the same casino, but the women tended to play games with a lower chance of success than the men. You can always increase your intrinsic probability of winning by playing the game with better chances. The correlation between being a woman and being rejected was an artifact of the composition of the ensemble of female applicants, in just the way that the higher relative probability of speaking English over Samoan is an artifact of the composition of the ensemble of persons, and in just the way that the higher relative probability of entropy increasing processes over entropy decreasing ones is an artifact of the composition of typical ensembles in worlds that start out in a state of low entropy.

---

[16] Assume, here, that births distribute themselves roughly evenly over the year in each country.

Another example of the way that an uneven partition can artificially favor some elements is that I am the most hyperlinked 'Jenann' on the internet. If one googles 'Jenann', links to me, or pages of mine, come up first, second, third, and fourth. But David Chalmers is not the most hyperlinked 'David' on the internet, and Saul Kripke is not the most hyperlinked 'Saul' on the internet. That does not mean that I am more famous than David Chalmers or Saul Kripke. The reason is that there are fewer Jenanns than there are Davids or Sauls (at least among the relevant population), and so I am more intrinsically likely to be most linked among my conomials. I am a big fish only relative to the size of my pond. The pond of my conomials is ever so much smaller than the pond of Davids or Sauls.

PRINCIPLE OF INDIFFERENCE

Some will have already spotted a connection with that notorious principle of probabilistic reasoning that has fallen into ill repute: the

Principle of Indifference (PI).[17] PI enjoins that given a set of possibilities and no specific reason for thinking one obtains rather than another, they should be assigned equal probability. It captures the intuition that, for example, if one has a six-headed die, one should assign probability 1/6 to each face on any given roll, or if one is presented with a spinning wheel labeled with numbers in the interval $0>x>1$, one should assign equal probabilities to obtaining a number greater than 0.5 and less than 0.5 on a random spin of the wheel. There is a wealth of counterexamples to PI that derive inconsistent results by applying the principle to the same events under different descriptions or considered as part of different partitions. Here is an especially simple one:[18] A factory produces cubes with side lengths up to 2 cm. What is the probability that a randomly chosen cube has side length under 1 cm? Apply PI to side length, one gets the answer 1/2. A factory produces cubes with face area up to 4 cm^2. What is the probability that a randomly chosen cube has face area under 1 cm^2? Apply PI to face area and one gets the answer 1/4. These are the same problem under different descriptions. One gets contradictory answers because application of PI to side length is tantamount to assuming that the factory produces cubes whose sides show a uniform distribution in length, whereas application of PI to face area is tantamount to assuming that the factory produces cubes whose faces show a uniform distribution in area. But a production process uniformly distributed in side lengths produces a different distribution of cube volumes than a process uniformly distributed in face areas.

The spread of cube volumes depends on whether one assumes uniformity over side length or face area, and which of these assumptions is correct is a question of fact. It is a matter of which partition is in fact even, determined by statistics for the cubes the factory actually turns out. PI will give the right results if *and only if* applied to the partition that actually divides the cubes coming out of the factory into

---

[17] This is also relevant to the Doomsday argument and Shooting Room set-up. In the Doomsday case, the natural metric imposed by supposing a person to be equally likely to live at any given time, and in the shooting room the metric imposed by regarding an arbitrary player to be equally likely to be playing in any given cycle of the game are what generate the mistaken expectations. Partitioning the population by date of birth, or partitioning players by game cycle, does not divide it into equally sized cells and expectations derived by projecting relative frequency in population overall is weighted in favor of later cells. An arbitrary member of the population is much more likely to live to see Doomsday and an arbitrary player of the shooting game is 9 times more likely to win than to die, though the chance of Doomsday at any one time is as likely as at any other, and the chance of losing in any cycle of the shooting game is just as likely as that of winning.

[18] Derived from Bas van Frasssen, *Laws and Symmetry* (New York: Oxford, 1989).

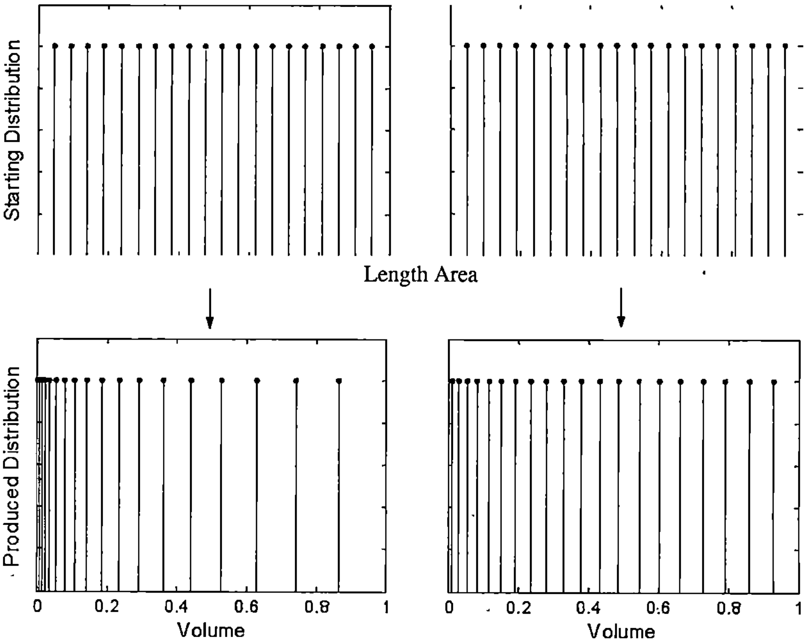


Figure 1: The top two panels represent uniformity in length and area. The bottom two panels show how these uniform distributions change when mapped to volume. A process uniform in length produces a different set of volumes than a process uniform in area.

equally sized classes. That is an unavoidably empirical question. There is no more a way of identifying the even partition from logical principles than there is of using such principles to divine what the factory heads decide at this year's corporate convention. Given a random spin of the wheel whose edges are labeled between 0 and 1, should one spread probability uniformly over equal intervals? Or should one use some other function? There are as many ways of spreading probability over the interval $0<x<1$ as there are real-valued functions whose arguments in that interval sum to 1. Which is correct depends on whether the wheel spins at uniform speed, and how numbers are actually distributed around the wheel. And, again, that is an unavoidably empirical matter. What is right for one wheel will be wrong for another. When PI is applied to reflect a lack of specific information about a system, it can be done in a well-defined manner only against the background of general information about how properties of interest distribute themselves across the population from which the system is drawn. Consider a circle of finite volume on which all colors in the visible spectrum are represented. If asked for the probability that a randomly chosen surface will have a color that falls on the top half of the wheel, many will automatically answer ½. This response is correct assuming the standard representation of the spectrum on which colors separated by equal wavelengths are separated by equal volumes on the circle. But it is perfectly possible to crowd all colors except for a particular shade of deep purple into the bottom half of the wheel. Nothing said when the question was introduced specified how colors were to be distributed around the circle, and the answer would be wrong on this representation. The ways families of properties are represented visually and imaginatively almost always incorporate assumptions, rarely explicit, about statistical probability, that is, the probability that a random pick from a typical ensemble of systems of the relevant type will have the property or properties of interest. It is these assumptions that are made explicit when their implications are drawn out in applications of PI. Richard Von Mises, discussing Bertand's chord paradox, says the same thing about the absence of formal criteria for identifying the right metric. By 'coordinate system', here, he means parameterization:

> ... the assumption of a 'uniform distribution' means something different in different co-ordinate systems. No general prescription for selecting 'correct' co-ordinates can be given, and there can therefore be no general preference for one of the many possible uniform distributions.[19]

Although one is rarely conscious that one is doing it, one chooses parameters that reflect one's metrical assumptions, and it is really the metrical assumptions built into the choice of parameters that underwrites the practice of assigning uniform distribution to reflect ignorance. The metrical assumptions are wholly empirical, and the practice is only successful to the degree that they are correct. Assigning probabilities to events to reflect a lack of specific information, that is, a lack of information that singles the system of interest out from the crowd, requires some general knowledge about the composition of the crowd. In physical contexts, this kind of general knowledge is embodied in the metrical structure of phase space. Specific information gets used to carve out a volume of phase space, but when specific information gives out, one gives a flat distribution over the remaining volume and general knowledge kicks in. If metrical structure faithfully reflects statistical probabilities, when it is interpreted by the principle that a random pick is equally likely to yield a system in any two equal volumes of the space, indifference over the volume of phase space whose boundaries reflect the extent of our specific knowledge will give us probabilities that reflect statistics in the relevant population.[20] We have to make *some* metrical assumption before we can form expectations, expectations derived from a flat distribution over an uneven partition or a space equal volumes of which do not represent statistically equiprobable states is going to leave us with distorted probabilities. In a finite example, there is an easy way of identifying the even partitions; we count the occupants of cells. In an infinite population, where proportions are replaced by measures and there is no intrinsic measure of size of classes, identifying the even partition requires all of the tools of modern statistics, and there are no logical guarantees that the statistics in any finite sample will be representative.

---

[19] Von Mises, *Probability, Statistics and Truth* (New York: Macmillan, 1957), pp. 80–81.

HIGHER ORDER IGNORANCE

Of course one can be ignorant of general matters too, and general ignorance will compound with lack of specific information when probabilities are calculated. General ignorance is ignorance of how properties of interest distribute themselves across the population whose boundaries reflect the extent of specific knowledge. It is represented by mixing probabilities drawn from different hypotheses about this distribution.[21] So, in the example above, if it is not known how the factory makes cubes, or it is not known how numbers are distributed around the spinning wheel, but one knows how to partition the alternatives into equiprobable classes, PI can be applied at a higher level by forming a mixture weighted to reflect higher order indifference between equiprobable alternatives. But one does not always know how to form an even partition of alternatives, and in some cases it does not make sense. This kind of higher order ignorance is complex. It remains, however, that the first order content of a physical theory, insofar as the theory is used to generate expectation in the face of ignorance, includes a probabilistic postulate which serves as part of the background against which the impact of incrementally obtained information is assessed. Phase spaces and state spaces are what statisticians refer to as 'sample spaces'. They embody a great deal of general knowledge of an empirical nature, and that knowledge is relied on when a theory is used to form expectation on the basis of limited specific information.[22]

---

[20] We have to be careful to understand that the indifference is not over possible outcomes for an experiment, but indifference over outcomes generated by the same method. Incorporate everything you know about the system into the process that generates the outcome, and think of selection as a random choice from outcomes generated by the same process. So, for example, if you are assigning probabilities to a dart thrown by John hitting the center circle, you take all of the specific knowledge you have of the process, and let the probabilities reflect the distribution of outcomes in that class (for example, darts thrown by John from a certain distance in bad light after he has had three beers, while he is talking to buddies, and so on).

[21] 'Mixing' is a technical notion, the mixture of probabilities P and P^* drawn from different sample spaces is $(\alpha P + \beta P^*)$ where α and β are the relative probabilities of the sample spaces from which they are drawn.

THE LAPLACIAN ARGUMENT

If what preceded is correct, there ought to be a gap in the knowledge of the Laplacian intelligence. There is. It is revealed by the fact that the Laplacian intelligence cannot answer problems of the form: 'What is the statistical probability with which a system in state A at t ends up in state B at t^*, given C?' where A and C together fall short of a complete specification of the microstate of the universe, and B is fine-grained enough not to permit a 1 or 0 value. By hypothesis, the Laplacian intelligence knows the microstate of the universe, and he knows the dynamical laws, and whenever he tries to make predictions about the behavior of a thermodynamic system, he evolves forward the microstate, so he never has to worry about probabilities and his predictions are never wrong. His extraordinary specific knowledge keeps him from confronting problems of this form. Problems of this form, however, are the rule for us, and the question is not whether the Laplacian intelligence confronts them, but whether he has what he needs to provide answers. The thermodynamic example demonstrated that he cannot squeeze answers out of the dynamical laws alone, for one can hold the laws fixed and generate radically different answers to questions of this form by adopting different probabilistic postulates. Looking at frequencies will provide *evidence* for probabilities, but as a logical matter they do not determine the probabilistic facts. Perhaps, as those that defend Humean reductions suppose, facts about probabilities are included implicitly in the manifold of categorical fact, not directly in the frequencies but along with the laws as part of the best overall systematization of the manifold.[23] That may be correct, but it is separate from the thesis that has been defended here: namely, that any systematization that can be used as a basis for prediction or receive confirmation from the evidence will incorporate a measure over phase space, and the measure will be just as real and objective as, but not reducible to, the physical laws.[24]

---

[22] The boundary between general and specific knowledge is conventional. What counts as specific knowledge about a particle in a context in which we are using a state space for generic particles (for example, that it is negatively charged) is trivial (has probability 1, for every particle) in a context in which we are using a state space for electrons.

[23] Loewer, "Determinism and Chance," *op. cit.*, and Lewis, "Humean Supervenience Debugged," *Mind*, CIII, 412 (1994): 473–90.

BAYESIANISM, AND THE BEST SYSTEMS ANALYSIS

A couple of residual remarks are warranted to draw connections to other work in the literature. Loewer's discussion of statistical mechanical probabilities in "Determinism and Chance" is in some ways in accord with the position that has been defended here. His argument begins with the so-called Best Systems Analysis of laws. According to the Best Systems Analysis, what makes a certain claim about the world a law is that it is part of a theoretical package that overall gets the best fit with the evidence. The idea is that the correct physical theory will be that systematization that achieves the best overall fit with the global pattern of local matters of particular fact. He argues on this basis that the microcanonical probability distribution associated with statistical mechanics has a claim to be included in a theoretical package alongside the laws because when it is added to the laws, it vastly increases the informativeness and fit of the best system. He writes

> If adding a probability distribution over initial conditions—in our examples over initial conditions of the universe—to other laws greatly increases informativeness (in the form of fit) with little sacrifice in simplicity then the system of those laws together with the probability distribution may well be the best systematization of the occurrent facts.[25]

I am in full accord that we should regard a probability postulate alongside the laws as part of the theoretical package. The differences between Loewer's position and mine are these: (i) I have argued that we should do this quite generally, not just in the context of statistical mechanics, (ii) I think we should take $\Pr_G(A/B)$ as the basic object, rather than a distribution over initial conditions. $\Pr_G(A/B)$, recall, is defined as the probability that a random pick from B-systems would yield an A. It can be generated from a distribution over initial conditions in the context of deterministic laws, but does not require the idea of an initial state and has applicability also in indeterministic contexts, (iii) I am resistant to regarding $\Pr_G(A/B)$ as a *law* for reasons that are independent of what was said here, and finally (iv) where Loewer thinks that including a probabilistic postulate increases fit with the evidence, I think that without a probabilistic postulate, we do not have an objective measure of fit with the evidence that goes beyond logical compatibility.[26]

---

[24] What our physical theories do is "solve" the data for generalizable patterns. I am arguing here that those generalizable patterns are not just the regularities embodied in physical laws, but patterns of migration over state space, not derivable from the laws without invoking a measure over state space, patterns that tell us how a group of systems, distributed across a finite volume of state space redistribute themselves over time.

[25] Loewer, "Determinism and Chance," *op. cit.*, p. 10.

Note that Bayesians can quite comfortably accept the thesis that a probabilistic postulate should be recognized as a standard component of a physical theory, even in deterministic contexts. One of the defining features of Bayesianism is the thesis that all probability is ultimately subjective, but there is some equivocation about what this actually entails. Subjective probability measures—in particular, the conditional probabilities that govern how belief is updated incrementally in response to evidence—encode a lot of information about the probabilistic relationships among events. A Bayesian can view a physical theory as nothing more than a candidate credence function, that is, a regimented rendering of the conditional connections among belief that govern the updating of opinion. This is what is known in the literature as an *expert function*. Insofar as one is viewing a physical in this way, she will have a special reason to recognize a probabilistic postulate as part of its content, for without a probabilistic postulate, a theory does not constitute a well-defined credence function.

CONCLUSION

Arguments have been given here for recognition of a type of probability that is more basic than chance, that is not intrinsically dynamical, and that is a practically indispensable component of deterministic and indeterministic theories alike: $\mathrm{Pr}_G(A/B) =_{\mathrm{def}}$(the probability that a random pick from B-systems will generate a system in A). When A and B are dynamical states, $\mathrm{Pr}_G(A/B)$ gives the transition probability from B to A. It also gives the transition probability from A to B, and the nondynamical probability that a system has one property, given that it has another, for example, the probability of being green, given the property of being an emerald or the property of being in microstate A, given the property of being in macrostate B. It encodes virtually all of the inductively important information in a physical theory, everything that cannot be strictly deduced from known fact *via* the laws. If one knows enough that the physical laws leave only one possibility, the measure is not needed. But absent perfect and precise knowledge, there is always an indefinite number of possibilities and the laws do not discriminate among them. They do not say which of a set of possibilities is *more or most* possible, and so they do not tell one how to divide opinion between them. As for the suggestion to divide opinion equally; 'dividing it equally' is not well defined until a metric has been specified. What is really meant when one is instructed to divide opinion equally is 'distribute it evenly *among equally probable alternatives*'. That just emphasizes the ineliminability of $\Pr_G$. If one is looking to predict where one does not have complete information, or to intervene where one does not have precise control, one needs probabilities, and the probabilities have to be objective in the sense that they have to reflect the statistics in the population of interest. The more detailed one's specific information about the system, and the better her knowledge of the statistics, the more likely her success.

---

[26] Any use of subjective probabilities will mean that agents with different priors will differ in their assessment of relative fit with the evidence.

In arguing that a probabilistic postulate should be made one of the standardly recognized components of a theoretical package, it is reasonable to ask whether this is intended as a redescription of assumptions implicit in practice or a recommendation to begin including a probabilistic postulate in statements of theory. The answer is that there is both an observation and a recommendation. The observation is that insofar as one derives expectations from theory in a real setting, one is in fact invoking probabilities. The recommendation is to bring the often tacit, intuitive measures employed in practice into the open by including them in the explicit content of a theory, where their role in generating expectations is made transparent, and where they can be subjected to scrutiny and compared with alternatives.

J.T. ISMAEL

Centre for Time, University of Sydney
and University of Arizona

NEW BOOKS: TRANSLATIONS

ADORNO, THEODOR W.: *Letters to His Parents.* WIELAND HOBAN, trans., CHRISTOPH GÖDDE and HENRI LONITZ, eds. Malden, MA: Polity, 2006. vii, 397 p. Cloth $45.00; *Towards a Theory of Musical Reproduction.* WIELAND HOBAN, trans., HENRI LONITZ, ed. Malden, MA: Polity, 2006. xxii, 278 p. Paper $24.95; *Dream Notes.* RODNEY LIVINGSTON, trans., CHRISTOPH GÖDDE and HENRI LONITZ, eds. Malden MA: Polity 2007. vii, 117 p. Cloth $17.95; *Lectures on Negative Dialectics: Fragments of a Lecture Course 1965/1966.* RODNEY LIVINGSTONE, trans. Malden, MA: Polity, 2008. xix, 267 p. Cloth $65.00, Paper $26.95.

AGAMBEN, GIORGIO: *Profanations.* JEFF FORT, trans. New York: Zone, 2007. 99 p. Cloth $25.95.

BEAUFRET, JEAN: *Dialogue with Heidegger: Greek Philosophy.* MARK SINCLAIR, trans. Bloomington: Indiana UP, 2006. xxvi, 154 p. Cloth $39.95.

BADIOU, ALAIN: *The Century.* ALBERTO TOSCANO, trans. Malden, MA: Polity, 2007. xiii, 233 p. Paper $22.95.

BAUDRILLARD, JEAN: *Cool Memories V.* CHRIS TURNER, trans. Malden, MA: Polity, 2006. 114 p. Paper $19.95.

BIDERMAN, SHLOMO: *Crossing Horizons: World, Self, and Language in Indian and Western Thought.* ORNAN ROTEM, trans. New York: Columbia, 2008. x, 356 p. Cloth $45.00.

BOFF, LEONARDO: *Essential Care: An Ethics of Human Nature.* ALEXANDRE GUILHERME, trans. Waco, TX: Baylor, 2007. xi, 178 p. Paper $29.95.

BRAGUE, RÉMI: *The Law of God: The Philosophical History of an Idea.* LYDIA G. COCHRANE, trans. Chicago: University Press, 2007. 365 p. Cloth $35.00.

CANGUILHEM, GEORGES: *Knowledge of Life.* PAOLA MARRATI and TODD MEYERS, trans. New York: Fordham, 2008. xx, 202 p. Cloth $75.00, paper $24.00.

CANTO-SPERBER, MONIQUE: *Moral Disquiet and Human Life.* SILVIA PAVEL, trans. Princeton: University Press, 2008. 222 p. Paper $39.50.

CAVARERO, ADRIANA: *Horrorism: Naming Contemporary Violence.* WILLIAM MCCUAIG, trans. New York: Columbia, 2009. viii, 154 p. Cloth $29.50.

CHERIF, MUSTAPHA: *Islam and the West: A Conversation with Jacques Derrida.* LAVENDER TERESA FAGAN, trans. Chicago: University Press, 2007. xxvii, 255 p. Paper $19.95.

CIXOUS, HÉLÈNE: *Dream I Tell You.* BEVERLEY BIE BRAHIC, trans. New York: Columbia, 2007. 143 p. Paper $18.95.

CONFUCIUS: *The Analects of Confucius.* BURTON WATSON, trans. New York: Columbia, 2007. 162 p. Cloth $19.95.

CSIKSZENTMIHALYI, MARK, ed. and trans.: *Readings in Han Chinese Thought.* Indianapolis: Hackett, 2006. xxx, 220 p. Cloth $39.95, paper $14.95.

DE BEAUVOIR, SIMONE: *Diary of a Philosophy Student, Volume I: 1926–27.* BARBARA KLAW, SYLVIE LE BON DE BEAUVOIR, MARGARET A. SIMMONS, and MARY BETH TIMMERMANN, eds., BARABAR KLAW, trans. Urbana: Illinois UP, 2006. xi, 374 p. Cloth $40.00; *Wartime Diary.* ANNE DEING CORDERO, trans. Urbana: Illinois UP, 2009. 350 p. Cloth $40.00.

DE SAUSSURE, FERDINAND: *Writings in General Linguistic.* CAROL SANDERS, MATTHEW PIRES, and PETER FIGUEROA, trans. New York: Oxford, 2006. xxx, 336 p. Cloth $55.00.

DERRIDA, JACQUES: *Geneses, Genealogies, Genres, and Genius: The Secrets of the Archive.* BEVERLY BIE BRAHIC, trans. New York: Columbia, 2006. xvi, 96 p. Cloth $24.50.

DÜTTMAN, ALEXANDER GARCÍA. *Philosophy of Exaggeration.* JAMES PHILLIPS, trans. New York: Continuum, 2007. 182 p. Cloth $110.00.

EBELING, FLORIAN: *The Secret History of Hermes Trismegistus: Hermeticism from Ancient to Modern Times.* DAVID LORTON, ed. and trans. Ithaca, NY: Cornell, 2007. xiii, 158 p. Cloth $29.95.

FOUCAULT, MICHEL: *History of Madness.* JONATHAN MURRAY and JEAN KHALFA, trans. New York: Routledge, 2007. xxxix, 725 p. Cloth $40.00.

GROEBNER, VALENTIN: *Who Are You: Identification, Deception, and Surveillance in Early Modern Europe.* MARK KYBURZ and JOHN PECK, trans. New York: Zone, 2007. 349 p. Cloth $30.00.

HADOT, PIERRE: *The Veil of Isis: An Essay on the History of the Idea of Nature.* MICHAEL CHASE, trans. Cambridge: Harvard, 2006. xii, 399 p. Cloth $29.95.

HEGEL, GEORG WILHELM FRIEDRICH: *Lectures on Logic: Berlin, 1831.* CLARK BUTLER, trans. Bloomington: Indiana UP, 2008. xxiii, 233p. Cloth $39.95.

HEIDEGGER, MARTIN: *Mindfulness.* PARVIS EMAD and THOMAS KALARY, trans. New York: Continuum, 2006. xliii, 385 p. Paper $29.95; *Basic Concepts of Ancient Philosophy.* RICHARD ROJCEWICZ, trans. Bloomington: Indiana UP, 2008. xiv, 253 p. Cloth $39.50; *Ontology: The Hermeneutics of Facticity.* JOHN VAN BUREN, trans. Bloomington: Indiana UP, 2008. xii, 138 p. Paper $19.95.

HEINE, HEINRICH: *On the History of Religion and Philosophy in Germany.* TERRY PINKARD, ed., HOWARD POLLACK-MILGATE, trans. New York: Cambridge, 2007. xli, 218 p. Cloth $85.00, paper $29.99.

HERACLITUS: *Heraclitus: Translations and Analysis.* DENNIS SWEET, ed. Lanham, MD: University Press of America, 2007. xiii, 96 p. Paper $19.95.

HOBBES, THOMAS: *Reason of State, Propaganda, and the Thirty Year's War: An Unknown Translation of Thomas Hobbes.* NOEL MALCOLM, trans. New York: Oxford, 2007. x, 227 p. Cloth $65.00.

HOLLWECK, THOMAS A., ed.: *Eric Voegelin: Selected Correspondence, 1950–1984.* SANDY ADLER, THOMAS A. HOLLIWECK, and WILLIAM PETROPULOUS, trans. Columbia: Missouri UP, 2007. 941 p. Cloth $74.95.

HONNETH, AXEL: *Disrespect: The Normative Foundations of Critical Theory.* Various translators. Malden, MA: Polity, 2007. xiii, 275 p. Paper $26.95.

JAMBET, CHRISTIAN: *The Act of Being: The Philosophy of Revelation in Mullā Sadrā.* JEFF FORT, trans. New York: Zone, 2006. 497 p. Cloth $38.95.

JULLIEN, FRANÇOIS: *The Impossible Nude: Chinese Art and Western Aesthetic.* MAEV DE LA GUARDIA, trans. Chicago: University Press, 2007. 136 p. Cloth $40.00; *Vital Nourishment: Departing from Happiness.* ARTHUR GOLDHAMMER, trans. New York: Zone, 2007. 176 p. Cloth $25.95.

KANT, IMMANUEL: *Toward Perpetual Peace and Other Writings on Politics, Peace, and History.* DAVID L. COLCLASURE, trans., PAULINE KLEINGLED, ed., contributions by JEREMY WALDRON, MICHAEL W. DOYLE, and ALLEN W. WOOD. New Haven: Yale, 2006. xxiv, 265 p. Cloth $45.00, paper $17.00; *Anthropology, History, and Education.* MARY GREGOR, et al., trans., GÜNTER ZÖLLER and ROBERT B. LOUDEN, eds. New York: Cambridge, 2008. xvi, 597 p. Cloth $150.00.

KIERKEGAARD, SØREN: *Fear and Trembling.* C. STEPHEN EVANS and SYLVIA WALSH, trans. New York: Cambridge, 2006. xxxvii, 115 p. Cloth $55.00, paper $16.99.

LACOUE-LABARTHE, PHILIPPE: *Heidegger and the Politics of Poetry.* JEFF FORT, trans. Urbana: Illinois UP, 2007. xviii, 111 p. Cloth $35.00.

LAÏDI, ZAKI: *The Great Disruption.* CHRIS TURNER, trans. Malden, MA: Polity, 2007. vii, 260 p. Paper $24.95.

LEIBNIZ, GOTTFRIED WILHELM: *Leibniz: The Shorter Leibniz Texts.* LLOYD STRICKLAND, trans. New York: Continuum, 2006. xx, 214 p. Paper $29.95.

LICHTENSTEIN, JACQUELINE: *The Blind Spot: An Essay on the Relations between Painting and Sculpture in the Modern Age.* CHRIS MILLER, trans. Los Angeles: Getty, 2008. ix, 215 p. Cloth $35.00.

LONGEWAY, JOHN LEE: *Demonstration and Scientific Knowledge in William of Ockham: A Translation of Summa Logicae III–II: De Syllogismo Demonstrativo, and Selections from the Prologue to the Ordinatio.* Notre Dame: University Press, 2007. xx, 432 p. Cloth $58.00.

LONGUENESSE, BÉATRICE: *Hegel's Critique of Metaphysics.* NICOLE J. SIMEK, trans. New York: Cambridge, 2007. xxi, 246 p. Cloth, $85.00.

MALABOU, CATHERINE: *What Should We Do with Our Brain?* SEBASTIAN RAND, trans. New York: Fordham, 2008. xv, 94 p. Cloth $55.00, paper $18.00.

MARRATI, PAOLA: *Gilles Deleuze: Cinema and Philosophy.* ALISA HARTZ, trans. Baltimore: Johns Hopkins, 2008. xix, 138 p. Cloth $45.00.

MARX, KARL: *The First Writings of Karl Marx.* PAUL M. SCHAFER, ed. Various translators. New York: Ig, 2006. 223 p. Paper $14.95.

MENGZI: *Mengzi: With Selections from Traditional Commentaries.* BRYAN W. VAN NORDEN, trans. Indianapolis, IN: Hackett, 2008. xlvii , 207 p. Cloth $37.95, Paper $12.95.

MICHELSTAEDTER, CARLO: *Persuasion and Rhetoric.* WILHELM SNYMAN and GIUSEPPE STELLARDI, trans. Scottsville, South Africa: University Press, 2007. xix, 153 p. Cloth $55.95.

MORO, ANDREA: *The Boundaries of Babel: The Brain and the Enigma of Impossible Languages.* IVANO CAPONIGRO and DANIEL B. KANE, trans. Cambridge: MIT, 2008. xvi, 257 p. Cloth $35.00.

NĀGĀRJUNA: *Reason Sixty: With Chandrakīrti's Commentary.* JOSEPH LOIZZO, trans. New York: Columbia, 2007. xxix, 434 p. Cloth $49.00.

NANCY, JEAN-LUC: *Listening.* CHARLOTTE MANDELL, trans. New York: Fordham, 2007. xii, 85 p. Cloth $65.00, paper $16.00; *Philosophical Chronicles.* FRANSON MANJALI, trans. New York: Fordham, 2008. xix, 72 p. Paper $16.00.

NOVALIS: *Notes for a Romantic Encyclopaedia: Das Allgemeine Brouillon.* DAVID WOOD, trans. Albany: SUNY, 2007. xxx, 290 p. Cloth $35.00.

PLATO: *The Symposium.* M.C. HOWATSON and FRISBEE C.C. SHEFFIELD, eds.; M.C. HOWATSON, trans. New York: Cambridge, 2008. xxxv, 91 p. Cloth $49.00, paper $13.99.

PROCLUS: *Commentary on Plato's Timaus: Volume III, Book 3, Part 1: Proclus on the World's Body.* DIRK BALTZLY, ed. and trans. New York: Cambridge, 2007. xii, 205 p. Cloth $85.00.

RICOEUR, PAUL: *Reflections on the Just.* DAVID PELLAUER, trans. Chicago: University Press, 2007. 265 p. Cloth $25.00.

RIEMEN, ROB: *Nobility of Spirit: A Forgotten Ideal.* MARJOLIJN DE JAGER, trans. New Haven: Yale, 2008. xl, 116 p. Cloth $22.00.

RORTY, RICHARD, and PASCAL ENGEL: *What's the Use of Truth.* PATRICK SAVIDAN, ed., WILLIAM MCCUAIG, trans. New York: Columbia, 2007. xii, 80 p. Cloth $12.95.

ROUSSEAU, JEAN-JACQUES: *Autobiographical. Scientific, Religious, Moral, and Literary Writings.* CHRISTOPHER KELLEY, trans. Hanover, NH: Dartmouth, 2006. xxvii, 332 p. Cloth $70.00; *Rousseau on Philosophy, Morality, and Religion.* Various translators. New York: Cambridge, 2007. xxx, 277 p. Paper $24.95.

SABUCO DE NANTES BARRERA, OLIVA: *New Philosophy of Human Nature.* MARY ELLEN WAITHE, MARIA COLOMER VINTRO, and C. ANGEL ZORITA, eds. and trans. Urbana: Illinois UP, 2007. 340 p. Cloth $50.00.

SARTE, JEAN-PAUL: *Existentialism Is a Humanism.* CAROL MACOMBER trans., JOHN KUKLA, ed. New Haven: Yale, 2007. xiv, 108 p. Cloth $9.95.

SCHELLING, F. W. J.: *Historical-critical Introduction to the Philosophy of Mythology.* MASON REICHLEY and MARKUS ZISSELSBERGER, trans. Albany: SUNY, 2006. xxiv, 240 p. Cloth $80.00; *Philosophical Investigations into the Essence of Human Freedom.* JEFF LOVE and JOHANNES SCHMIDT, trans. Albany: SUNY, 2006. xxxv, 183 p. Cloth $50.00.

SERRES, MICHEL: *The Parasite.* LAWRENCE R. SCHEHR, trans. Minneapolis: Minnesota UP, 2007. xxvii, 255 p. Paper $19.95.

SVENDSEN, LARS: *Fashion: A Philosophy.* JOHN IRONS, trans. London: Reaktion, 2006. 188 p. Cloth $24.95; *A Philosophy of Fear.* JOHN IRONS, trans. London: Reaktion, 2008. 114 p. Cloth $19.90.

SZENDY, PETER: *Listen: A History of Our Ears.* CHARLOTTE MANDELL, trans. New York: Fordham, 2008. xiii, 160 p. Cloth $75.00, paper $22.00.

THELLE, NOTTO R.: *Seeking God's Face: Faith in an Age of Perplexity.* BRIAN MCNEIL, trans. Mahwah, NJ: Paulist, 2008. vi, 144 p. Cloth $14.95.

UNAMUNO, MIGUEL DE: *Treatise on Love of God.* NELSON R. ORRINGER, trans. Urbana: Illinois UP, 2006. xxxviii, 180 p. Cloth $40.00.

XENOPHON: *The Anabasis of Cyrus.* WAYNE AMBLER, trans. Ithaca: Cornell, 2007. xv, 281 p. Cloth 49.95, paper $16.95.

YANNARAS, CHRISTOS: *Person and Eros.* NORMAN RUSSELL, trans. Brookline, MA: Holy Cross Orthodox, 2008. xix, 395 p. Paper $24.95.

ZARADER, MARLÈNE: *The Unthought Debt: Heidegger and the Hebraic Heritage.* BETTINGA BERGO, trans. Stanford: University Press, 2006. 255 p. Cloth $65.00, paper $24.95.

THE JOURNAL OF PHILOSOPHY

Special Issues Available

Parts and Wholes — December 2006
On Bernard Williams — June 2003
Causation — April 2000
Jürgen Habermas and John Rawls — March 1995

The Dewey Lectures:
Michael Dummett — *Truth and the Past*, 2002
Hilary Putnam — *Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind*, 1994
Donald Davidson — *The Structure and Nature of Truth*, 1990
Amartya Sen — *Well-being, Agency, and Freedom*, 1985
John Rawls — *Kantian Constructivism in Moral Theory*, 1980
Ernest Nagel — *Teleology Revisited*, 1977
W.V. Quine — *Ontological Relativity*, 1968

The Woodbridge Lectures:
John McDowell - *Having the World in View: Sellars, Kant, and Intentionality*, 1997

Subscriptions

Individuals $45.00
Students $20.00
Libraries and Institutions $100.00
Print and Online for Libraries and Institutions $180.00
Postage outside the United States $15.00

Special Offer

One complimentary issue of any of the above issues with each new subscription.
All issues are $5.00 plus $3.00 for shipping and handling.
Send check or money order to:

The Journal of Philosophy
Columbia University
1150 Amsterdam Avenue, Mail Code 4972
New York, NY 10027

www.journalofphilosophy.org

REVUE
INTERNATIONALE DE
PHILOSOPHIE

Direction: Michel Meyer

c / o Université Libre de Bruxelles, C.P. 188
50, avenue Franklin-Roosevelt / B-1050 Bruxelles (Belgique)
Tél. 32 2 650 25 61 / Fax 32 2 650 45 61
e-mail : tsalis@ulb.ac.be

1-2009 / volume 63 / n° 247

Shakespeare

Agnes HELLER
Preface

Paul KOTTMAN
Disinheriting the Globe: On Hamlet's Fate

Peter MURPHY
The Power and the Imagination:
The Enigmatic State in Shakespeare's English History Plays

David ROBERTS
Likeness and Difference: The Logic of Plot in *Measure for Measure*

Zoltán GYENGE
The Heteronomy of sin. Richard, Agamemnon or perhaps Abraham?

Prochains fascicules / Forthcoming

Philosophie du roman
Goethe
Hintikka
Le théâtre des Grecs au romantisme

Graduate Faculty Philosophy Journal

THE NEW SCHOOL
A UNIVERSITY

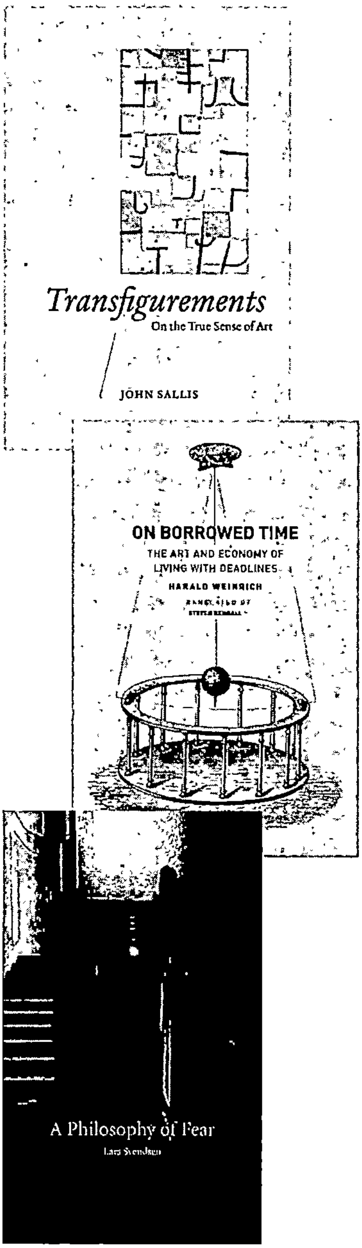
The **Graduate Faculty Philosophy Journal**, published biannualy in association with the Department of Philosophy, New School for Social Research, is a forum for historical and contemporary issues in philosophy. The *Journal* regularly publishes special issues dedicated to investigating the historical development of concepts and problems in the history of philosophy and their ongoing relevance to issues in contemporary philosophy.

Issue 30:1 April 2009

Merleau-Ponty at 100

Étienne Bimbenet	*Merleau-Ponty and the Quarrel over the Conceptual Contents of Perception*
Mauro Carbonne	*The Sensible Ideas Between Life and Philosophy*
Pierre Cassou-Nogués	*Merleau-Ponty, Whitehead, and Russell on Monadology and the Problem of Particulars*
Bernard Flynn	*Merleau-Ponty and Lévi-Strauss as Critics of Sartre*
Michael Gendre	*Nature and its Supplements in Merleau-Ponty's Collège de France* La Nature *Lecture Course*
Leonard Lawlor	*The Friend of the Future: A Note on Merleau-Ponty's "Brouillion d'une rédaction"*
Emmanuel de St. Aubert	*Space and the Image of the Body in Merleau-Ponty's Philosophy of the Flesh*
Ted Toadvine	*Truth and Resistance*

All communications should be addressed to the Editor, **Graduate Faculty Philosophy Journal**, Department of Philosophy, New School for Social Research, 79 Fifth Avenue Suite 1028, New York, NY 10003, or by email: gfpj@newschool.edu Domestic rates: Individuals: $25.00/year; Students: $15.00/year; Institutions: $50.00/year. Libraries and institutions can access *GFPJ* online by subscribing to the *Poesis* database. For more information, please contact order@pdc.net, or visit our website, www.newschool.edu/nssr/gfpj.



PHILOSOPHY

from Chicago

TRANSFIGUREMENTS
On the True Sense of Art
John Sallis

"With great rigor and uncommon delicacy, Sallis traces the sense—in all of its particularity—opened by art. The prose here is clear and the argumentation is compelling. Sallis bears witness to that which we tend to pass over far too soon; his capacity to linger at the site of the sensible makes reading this book a real pleasure."—Dennis J. Schmidt, Penn State University

CLOTH $25.00

ON BORROWED TIME
The Art and Economy of Living with Deadlines
Harald Weinrich
Translated by Steven Rendall

"Harald Weinrich is a phenomenon whose writing is nearly unique for its grace and accessibility. *On Borrowed Time* stands out among his many books for its totally unexpected topic. The more you think about it, the more you contemplate the extraordinary reach and unerring precision of the materials, the more fascinating it grows."—Marshall Brown, University of Washington

CLOTH $32.00

New from Reaktion Books

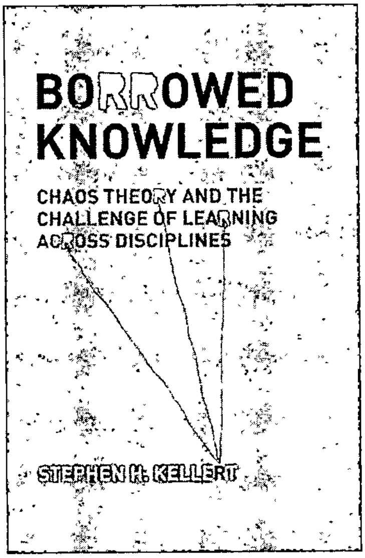
A PHILOSOPHY OF FEAR
Lars Svendsen

"In this brief yet wide-ranging and insightful book, Svendsen argues convincingly that we need to replace the risk society with a culture of hope and trust."—*Guardian* (UK)

PAPER $27.95

THE UNIVERSITY OF CHICAGO PRESS • www.press.uchicago.edu

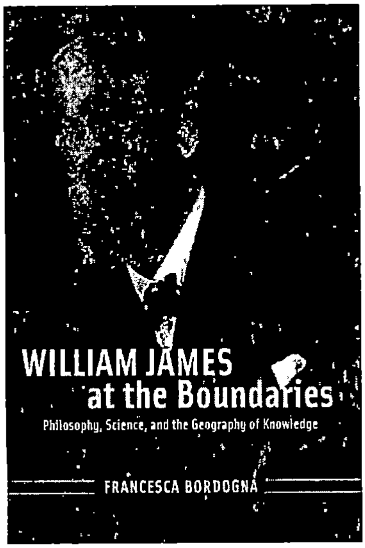
PHILOSOPHY OF SCIENCE *from Chicago*


BORROWED KNOWLEDGE
CHAOS THEORY AND THE CHALLENGE OF LEARNING ACROSS DISCIPLINES
STEPHEN H. KELLERT

BORROWED KNOWLEDGE
Chaos Theory and the Challenge of Learning across Disciplines
Stephen H. Kellert

"Stephen Kellert combines philosophy with rhetoric in a critical engagement with economists, literary, and legal theorists who import chaos theory in an attempt to widen and deepen their studies. His book is a model in two senses: it shows us how to borrow knowledge responsibly and how to criticize others responsibly when they borrow knowledge from other disciplines."—Alan G. Gross, University of Minnesota

CLOTH $35.00


WILLIAM JAMES at the Boundaries
Philosophy, Science, and the Geography of Knowledge
FRANCESCA BORDOGNA

WILLIAM JAMES AT THE BOUNDARIES
Philosophy, Science, and the Geography of Knowledge
Francesca Bordogna

"By placing James squarely at the creative intersection of the various disciplines, professions, and social activities of his time, Bordogna challenges the newly rigidified disciplinary habits that we instead inherited. This thought-provoking work makes an important contribution to the history of science and psychology."—Charlene Haddock Seigfried, Purdue University

CLOTH $39.00

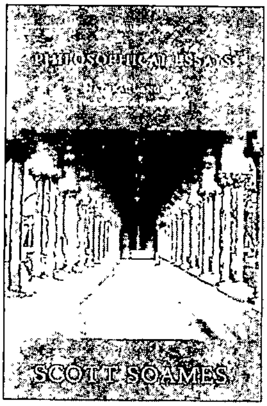
New from CSLI PUBLICATIONS

REASONING, RATIONALITY AND PROBABILITY
Edited by Maria Carla Galavotti, Roberto Scazzieri, and Patrick Suppes

Drawing on contributors across disciplines including philosophy, economics, psychology, statistics, computer science, engineering, and physics, *Reasoning, Rationality, and Probability* argues that the search for strong theories should leave room for the construction of context-sensitive conceptual tools.

PAPER $30.00

THE UNIVERSITY OF CHICAGO PRESS • www.press.uchicago.edu



Philosophical Essays, Volume 1

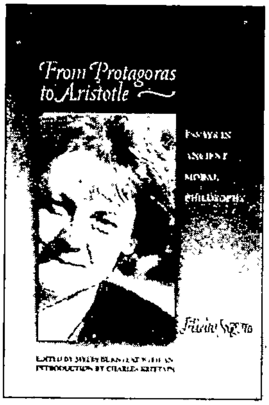
Natural Language: What It Means and How We Use It

Scott Soames

These volumes bring together the most important essays written by one of the world's foremost philosophers of language. A judicious collection of old and new, these volumes include sixteen essays published in the 1980s and 1990s, nine published since 2000, and six new essays.

The essays in *Volume 1* investigate what linguistic meaning is; how the meaning of a sentence is related to the use we make of it; what we should expect from empirical theories of the meaning of the languages we speak; and how a sound theoretical grasp of the intricate relationship between meaning and use can improve the interpretation of legal texts.

Paper $39.50 978-0-691-13681-3
Cloth $90.00 978-0-691-13680-6



From Protagoras to Aristotle

Essays in Ancient Moral Philosophy

Heda Segvic

Edited by Myles Burnyeat
with an introduction by Charles Brittain

This is a collection of the late Heda Segvic's papers in ancient moral philosophy. The papers range from a literary study of Homer's influence on Plato's Protagoras to analytic studies of Aristotle's metaphysics and his ideas about deliberation. Most of the papers reflect directly or indirectly Segvic's idea that both Socrates' and Aristotle's universalism and objectivism in ethics could be traced back to their opposition to Protagorean relativism. The book represents the considerable achievements of one of the most talented scholars of ancient philosophy of her generation.

Cloth $45.00 978-0-691-13123-8

PRINCETON UNIVERSITY PRESS
800.777.4726
press.princeton.edu

Datta
1-7-09
CUTTA
PORE CA

THE JOURNAL OF PHILOSOPHY

VOLUME CVI, NUMBER 3
MARCH 2009

page
113 *Being and Being Called: Paradigm Case Arguments and Natural Kind Words* Diego Marconi
137 *The Four Causes* Boris Hennig

COMMENTS AND CRITICISM
161 *Pains, Imperatives, and Intentionalism* Maura Tumulty

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Michael J. Mooney; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2009

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2009 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 3, MARCH 2009

BEING AND BEING CALLED: PARADIGM CASE ARGUMENTS AND NATURAL KIND WORDS*

What is the relation between being a *P* and being called '*P*', for example, between being a cat and being called 'cat'? Surely something might be a cat without being called 'cat'; indeed, cats as such might not be called 'cats'. If the word 'cat' disappeared from the language, the event would not entail the disappearance of cats. What about the converse implication? Does being called 'cat' entail being a cat? It would seem so. For suppose 'cat' refers to certain objects, and let Moon be one such object. Consider the statement that Moon is a cat. The statement is true just in case Moon is one of the objects that 'cat' refers to, which she is. Hence, the statement is true, therefore Moon is a cat. Being among the objects that 'cat' refers to entails being a cat.

However, it is one thing to say that 'cat' *refers* to certain objects, and (possibly) a different thing to say that certain objects *are called* 'cats'. Spiders are often called 'insects', yet 'insect' does not refer to spiders: it is not *correct* to call spiders 'insects'. "Being called" is often intended as a descriptive notion: whether something is or is not called '*P*' is just a fact that can be stated in terms of people's behavior or patterns of behavior. Reference, on the other hand, may not be descriptive in this sense. Philosophers have often been trying to characterize the quasi-technical notion of reference by suitably qualifying the everyday notion of "being called." Success in such an enterprise would amount to showing that being called* '*P*'—a suitably modified version of being called '*P*'—entails being *P*. Whether the enterprise is bound to fail is not the topic of this article. Here, I would like to show that one such attempt did fail, whereas another, more recent attempt that would

---

*Thanks to Andrea Bianchi, Paolo Casalegno, Andrea Iacona, Mark Sainsbury, and Achille Varzi for comments on previous versions.

0022-362X/09/0603/113-36

seem to be bound to fail for analogous reasons does not fail; or not for such reasons, anyway.

A few decades ago, some philosophers believed that being called '*P*' was (with some qualifications) a sufficient condition for being a *P*. They were the proponents of so called "paradigm case" arguments (J.O. Urmson, Antony Flew, Norman Malcolm, Max Black, and others; Ludwig Wittgenstein may have been among them). They argued that if something is *paradigmatically* called 'cat'—that is, if it is uncontroversial that anybody in his right mind and in full command of English would call it 'cat'—then it *is* a cat. In such circumstances, it would be foolish, or semantically deviant to deny that there are cats. Thus a skeptic about cats would be *ipso facto* refuted by exhibiting a paradigmatic application of the word 'cat'.

Paradigm case arguments went out of fashion in the late 1960s, not necessarily for very good reasons. However, I believe, and will try to show in what follows that such arguments were indeed flawed. The reason they were flawed is that being paradigmatically called '*P*' may not be sufficient for being *correctly* called '*P*'. But it is being correctly called '*P*', not just being called '*P*' or even being paradigmatically called '*P*', that entails being a *P*.

More recently, claims were put forward implying that, for '*P*', a natural kind term, paradigmatic *P*s could not fail to be *P*s, though in a partly different sense of 'paradigmatic'. It was claimed that being a cat is having the same nature as *these* things we usually call 'cats'.[1] Now, some of the things we *de facto* call 'cats' may be anomalous by not having the same nature as most things we call 'cats': they may be fakes, or perhaps pseudo-cats really belonging to a different natural species. To allow for such anomalies, we shall say that being a cat is having the same nature as *most* of the things we usually call 'cats'. The condition is taken to be both necessary and sufficient. By contrast, features that are usually associated with being a cat, such as being a mammal, chasing mice, and so on, are neither necessary nor jointly sufficient for being a cat. For example, if we were to discover that the objects we call 'cats' are really robots, not animals, we would persist in calling them 'cats'; and rightly so, for they *are* cats.[2] Being a living being is not a necessary condition for being a cat, or for being called 'cat'. Moreover, if in such circumstances an animal were discovered in the Equatorial jungle that possessed just the features we now attribute to cats —that is, being a feline, meowing, and so on—it would not qualify as a cat (we would say: This animal is exactly what we had taken cats to be!).

---

[1] Hilary Putnam, "The Meaning of 'Meaning'," in *Mind, Language, and Reality: Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 215–71.

[2] Putnam, "Is Semantics Possible?" in Howard E. Kiefer and Milton K. Munitz, eds., *Languages, Belief, and Metaphysics* (Albany: SUNY, 1970), pp. 50–63; later in *Mind, Language, and Reality: Philosophical Papers*, Volume 2, pp. 139–52. As Saul Kripke pointed out, if the world is the way we take it to be, that is, populated with mammals of a certain shape, size, and DNA (the animals we usually call 'cats'), then we *could not* discover that such mammals are robots, for they *necessarily* possess some among such features:

On this view, just being called 'cat' is neither necessary nor sufficient for being a cat. Something may be a cat even though the label 'cat' was never applied to it by anybody; conversely, something may be called 'a cat' by mistake, while not being a cat at all. However, that some beings rather than others are customarily called 'cats' is crucial for being a cat, for such a linguistic custom fixes the sample, so to speak, the nature of whose members (*if* they share a nature) eventually determines what being a cat is. Thus we may say that, on the view we are discussing, being called '*P*' entails being a *P* collectively, though not individually: the *P*s are the beings that are called '*P*' (excepting a few anomalies) and whatever other beings share their nature. Emphatically, the *P*s are *not* the beings that fit some definition of '*P*', or possess the features that are "analytically" associated with the word '*P*'.

We could then say—coming back to reference—that on this view a natural kind word such as 'cat' refers to *x* if and only if *x* shares a common nature with most of the beings we customarily call 'cats'. Is this form of the reduction of the notion of reference to the everyday notion of "being called" successful? At first sight, it seems that it is liable to the same kind of objection that can be raised against arguments based on paradigms: in some cases, it may not be correct to call *x* '*P*' even though *x* shares the same nature as most beings we call '*P*', for there are further constraints on the correct use of '*P*'. We shall see, however, that the view we are discussing—that is, the standard view about the reference of natural kind words—can be defended against such objections, at some cost.

I. PARADIGM CASE ARGUMENTS

If something, *o*, is a *P*, for example, a cat, or a past event, or an act of free will, it is correct to call it '*P*', when speaking English. An individual speaker may not be in a position to describe it as '*P*', perhaps because she misidentifies it or does not identify it at all, or does not have the word '*P*' in her lexical repertory; however, even for such a speaker

---

for example, there is no possible world in which cats are not living beings (see section IV below). However, disastrous mistakes cannot be ruled out, least of all by a theory of reference: the world might not be the way we take it to be (a different world might be actual). In that sense, we might discover that the objects we call 'cats' are really robots.

it would be correct—though perhaps unjustified—to call *o* '*P*'. It is conceivable that, in a whole linguistic community, *nobody* is in a position to call *o* a *P*, for any combination of the above reasons: even in such circumstances it would be correct, though presumably ungrounded, for each member of the community to apply the word '*P*' to *o*. For saying that it is not correct to call *o* '*P*' amounts to saying that *o* is not a *P*.

Conversely, if *o* is not a *P* it is not correct to call it '*P*' in English. *o* may look a lot like a *P*, in which case it is excusable that someone may call it '*P*'. Whether or not *o* resembles genuine *P*s, it may have become customary to call it a '*P*', so that it is natural for everybody in the community to call it '*P*' (it would be somewhat peculiar *not* to call it '*P*'). Or again, *o* may be indistinguishable from genuine *P*s given our cognitive resources, in which case it is inevitable for everybody to call it '*P*'. Nevertheless, in none of these cases is it correct to call it '*P*', no matter how excusable, natural, or even inevitable that it be called '*P*'. If *o* is not a *P*, it cannot be correct to call it '*P*'.

Thus, it appears that the following holds unconditionally:

(A) For every *x*, *x* is a *P* if and only if it is correct to call *x* '*P*'.

(A) has been called a "meaning platitude."[3] In fact, it is less platitudinous than it seems. If 'correct' is regarded as involving normativity, then (A) appears to allow for the derivation of an *ought* from an *is*. Moreover, people who are reluctant to Twin-Earth intuitions will doubt that, even if *o* is indistinguishable from *P*s given our (collective) cognitive resources, still it is not *correct* to call it '*P*'. They will argue that correctness in the use of language involves reference to some available standard: no linguistic behavior can be incorrect that is not in violation of some communal standard. But nothing counts as a violation of a standard if it cannot be pointed out as such by some member of the linguistic community: therefore, calling *o* a *P* cannot be said to be incorrect, even if *o* is not a *P* (in the eyes of God or of present-day science). However, for the purposes of the present discussion I will disregard all such difficulties with principle (A), which I will take to hold without restrictions.

Now consider the following proposal. Suppose that *o* is regularly described as a *P*, and that the disposition to apply '*P*' to *o* and the likes of *o* is regarded as an integral part of semantic competence relative to '*P*', so that of a speaker who is not prepared to call *o* '*P*' we would say that he does not know what '*P*' means. For example, if the speaker is puzzled by the fact that *o* is being described as 'a *P*', or is in trouble obeying orders or answering questions involving '*P*' even though *o* is present and circumstances are optimal for identification, we would say he does not know what '*P*' means.[4] Call the application of '*P*' to *o* (under these conditions) a *paradigm case* of the use of '*P*'; and call *o* a paradigm of *P* (or, synonymously, a paradigmatic *P*).[5] That is, assume the following definition:

---

[3] Anandi Hattiangadi, "Is Meaning Normative?" *Mind &Language*, XXI (2006): 220–40, p. 222.

> (For every *x*) the application of '*P*' to *x* is a paradigm case of the use of '*P*' if (and only if)
>
> (i) *x* is regularly called '*P*',
> (ii) being disposed to call *x* '*P*' is a necessary condition of semantic competence concerning '*P*'.

Could a paradigm case of the use of '*P*' fail to be a correct use of '*P*'? In other words, could it fail to be correct to describe *o* as a *P*, if the application of '*P*' to *o* is a paradigm case of the use of '*P*'? Could (B) fail?

> (B) For every *x*, if *x* is paradigmatically called '*P*' then it is correct to call *x* '*P*'.

If not, then if *o* is paradigmatically called '*P*' it cannot fail to *be* a *P*, by (B) and (A). So, if some skeptic were to deny that there are *P*s it would be enough to exhibit a paradigm case of the use of '*P*' in order to refute him.

This, or something very much like this seems to be the conceptual apparatus grounding so-called *paradigm case arguments*. The general pattern was the following. Suppose someone, *M*, claims that there are no *P*s. Now, here is an object *o* (an event *e*) that we would ordinarily describe as '*P*' or 'a *P*'; indeed, application of '*P*' to *o* (to *e*) is *a paradigm case* of the use of '*P*'. Therefore, *o* is *P* (*e* is *P*); consequently, either *M*'s claim is false, or she is using '*P*' in a nonstandard sense, so that her claim is not to the effect that there are no *P*s in the *ordinary* sense of '*P*'.[6]

Arguments of this form are intended to reverse the burden of proof. Against the skeptic who challenges—say—the reality of the past, one does not argue for the past's reality; instead, one shows that, given the ordinary meaning of the word 'past', the skeptic's claim is either plainly false (for some events are obviously past events) or semantically deviant. It then falls upon the skeptic to show that, in spite of the fact that 'past' paradigmatically applies to certain events, it should not be so applied, or, alternatively, that there are no such events. The burden of proof has been reversed.

---

[4] Max Black, "Making Something Happen," in Sidney Hook, ed., *Determinism and Freedom in the Age of Modern Science* (New York: Collier, 1958), pp. 31–45, on p. 33.

[5] For a clear presentation of the paradigm case argument and a survey of objections see Keith S. Donnellan, "Paradigm-case Argument," in Paul Edwards, ed., *The Encyclopedia of Philosophy* (New York: Macmillan, 1967), Volume 6, pp. 39–44. Oswald Hanfling, *Philosophy and Ordinary Language* (New York: Routledge, 2000), chapter 5, is a recent attempt at rescuing the paradigm case argument from standard objections.

[6] See Edwards, "Bertrand Russell's Doubts about Induction," in Anthony Flew, ed., *Logic and Language, First Series* (Oxford: Basil Blackwell, 1951), pp. 55–79, p. 63, for a formulation very close to mine.

The following are real-life examples (in schematic form) of paradigm case arguments:

Example 1:

(1) Suppose someone, *M*, claims that only present experience is real: the past is unreal.
(2) Nevertheless, I had breakfast this morning: this is what we call 'a past event'.
(3) Indeed, it is part of the grammar of the word 'past' that this is what we call 'a past event': application of 'past' to my having had breakfast this morning is a paradigm case of the use of 'past'.
(4) Therefore, my having had breakfast this morning is a past event.
(5) Consequently, *M* must be using 'past' (and 'present') in some nonstandard sense.[7]

Example 2:

(1) Suppose someone, *M*, denies that there are causes, holding that the notion of cause is self-contradictory.
(2) Consider a thirsty man who stretches out his hand, brings a glass of beer to his lips, and drinks. Such a sequence of acts is what we call 'making something happen'.
(3) Indeed, the act of reaching for a glass and drinking is an example of something that is used as a *standard* for the correctness of application of 'making something happen'.
(4) Therefore, the thirsty man made something happen. But to make something happen is to cause something to happen.
(5) Consequently, *M* must be using 'cause' in some nonstandard sense.[8]

In a paradigm case argument, one premise (premise 2 in both the above examples) presents an example of something that is uncontroversially described as 'a *P*': for example, a cause, or a past event. Both conjuncts of the premise must be satisfied: the existence of object *o*, or the occurrence of event *e* must be uncontroversial, and it must be plain that *o* (or *e*) would be normally described as '*P*'. Thus it is not enough, to refute skepticism about miracles,[9] that the turning of water into wine would be ordinarily described as a miracle, for it is far from uncontroversial that such an event ever took place; and it does not suffice that it be granted that *some* event actually took place at Cana of Galilea, for it is not plain that that event would be, or would have been described either as the turning of water into wine or as a miracle. Surely the word 'miracle' partly draws its meaning from its connection with alleged events such as one of water turning into wine; however, *alleged* events do not suffice for a paradigm case argument to get off the ground.[10]

---

[7] Based on Wittgenstein, *Philosophische Bemerkungen*, Rush Rhees, ed. (Oxford: Blackwell, 1964), §54.

[8] Based on Black, "Making Something Happen."

On the other hand, that the paradigms' existence must be uncontroversial does not entail that they have to be demonstrated. They can just as well be linguistically described, as long as it is not disputed that there are objects or events that fit the description. Nothing essential is gained by materially exhibiting a paradigm, for (as we shall see in a moment) the skeptic can still claim that what has been exhibited is not really a *P*, whether or not it is called '*P*'. On the other hand, nothing is lost by introducing paradigms by way of linguistic description, for the objection that nothing fits the description is not essentially stronger than the objection that the paradigms, though existing, are not really *P*s.[11]

Against the viability of paradigm case arguments in general, it was objected that the existence of objects or the occurrence of events that are paradigmatically called '*P*' is no proof that they are (genuine) *P*s: the skeptic could argue that they only *appear* to be *P*s.[12] By such a claim, the skeptic would acknowledge the meaning-constitutive connection between the word '*P*' and the alleged paradigms, while sticking to her belief that there really are no *P*s (only things that *look like P*s). This objection was countered with the remark that assertions like "This appears to be a *P*" are parasitic upon assertions like "This is a *P*."[13] I take this to mean that, for an assertion of the form "This appears to be a *P*" to be meaningful, some assertion of the form "This is a *P*" must be true: the possibility of apparent *P*s cannot be admitted without admitting the existence of genuine *P*s. But this seems too strong. Couldn't we say of some old woman that she looks like a witch without thereby committing ourselves to the existence of *bona fide* witches? In the general case, it seems that for an assertion of the form "This looks like a *P*" or "This appears to be a *P*" to make sense it is only required that there is some idea of what it is to look like a *P*: thus we can say (perhaps truly), of a very peculiar donkey, that it appears to be a unicorn (because of the animal's frontal protuberance, funny beard, and so on) while hastening to add that, of course, it *only appears* to be a unicorn—in fact, unicorns do not exist.

---

[9] J.W.N. Watkins, "Farewell to the Paradigm-Case Argument," *Analysis*, XVIII (1957): 25–33, see pp. 26–27.

[10] Cf. Scott Soames, "Malcolm's Paradigm Case Argument," chapter 7 of *Philosophical Analysis in the Twentieth Century*, Volume 2: *The Age of Meaning* (Princeton: University Press, 2003), pp. 157–70, see p. 162.

[11] Cf. Hanfling, *Philosophy and Ordinary Language*, p. 76.

[12] Robert Richman, "On the Argument of the Paradigm Case," *Australasian Journal of Philosophy*, XXXIX (1961): 75–81, see p. 77; the objection is endorsed by Donnellan, "Paradigm-Case Argument," p. 42.

[13] Christopher John Fards Williams, "More on the Argument of the Paradigm Case," *Australasian Journal of Philosophy*, XXXIX (1961): 276–78.

Suppose, then, that the skeptic is pushing the appearance point; and suppose that she is a skeptic about elephants (as in Keith Donnellan, *op. cit.*). She is claiming that paradigmatic elephants only *appear* to be elephants; there really are no elephants. Given the generally uncontroversial existence of paradigmatic elephants, what could she mean? One thinks of visual illusions, on the one hand, of counterfeits on the other: all uses of 'elephant', including paradigmatic uses, are occasioned by mirages or holograms; or else, all "elephants" are sophisticated robots made of plastic and metal. Now, suppose the skeptic is right: on inspection, more and more elephants turn out to be robots (including the familiar elephant at the local zoo). What would this prove exactly? Would such a finding warrant the conclusion that there really are no elephants? After all, the friend of paradigms never claimed that elephants are animals (not robots); only that *there are* elephants.

II. THE REFERENTIAL/INFERENTIAL DIALECTIC

Here we come to one of the central points in the present discussion. When the skeptic claims that there really are no elephants (only things that appear to be elephants), she is not denying that certain objects are customarily called 'elephant'; she may even grant that there are paradigmatic uses of 'elephant', that is, that the ability to call certain large, grey objects 'an elephant' is regarded as part of semantic competence relative to the word 'elephant'. Neither is she questioning that such objects really exist.[14] What, then, is she questioning? Apparently, that those things really are elephants. On what grounds? If she regards the discovery that the alleged elephants are made of plastic and metal as conclusive evidence for her claim, she must be assuming that elephants could not be robots: they are supposed to be living beings, animals, Probocideans, and so forth. For if she were not relying on some such assumption, she could only claim to have discovered that *elephants* have turned out to be robots, which surely does not entail that elephants do not exist (or that *those* things are not really elephants). In other words, the skeptic is assuming that there are conditions for being an elephant over and beyond being called 'elephant', albeit paradigmatically: being a living being is one such condition, which robots do not meet. The skeptic is assuming what I shall call a *notion* of elephant. While the friend of paradigms is convinced that being paradigmatically called '*P*' is sufficient for being a *P*, the skeptic insists that that is not enough (indeed, it may be regarded as irrelevant): in order to be a *P* it is necessary to have properties Q_1, Q_2,..., Q_n. The friend of paradigms is, as I shall say, a *referentialist* about being *P*: to be *P*, it is enough to be something to which '*P*' paradigmatically applies. By contrast, the skeptic is an *inferentialist*: being a *P* entails being a Q_1, and being a Q_2,..., and being a Q_n.[15] The conflict between the skeptic and the friend of paradigms is about whether being *P* must be conceived referentially or inferentially.

---

[14] What about the skeptic who claims that there are no elephants, only hallucinations of elephants? Even she may not deny that such hallucinations occur in typical circumstances, and that being able to use the word 'elephant' in such circumstances is part of semantic competence relative to the word. On the other hand, some philosophers had the skeptic concede that no hallucinations are involved (Norman Malcolm, "Moore and Ordinary Language," in P.A. Schilpp, ed., *The Philosophy of G.E. Moore* (Evanston, IL: Open Court, 1942), pp. 343–68; later in V.C. Chappell, ed., *Ordinary Language* (Englewood Cliffs, NJ: Prentice-Hall, 1964), pp. 5–23; J.N. Findlay, "Time: A Treatment of Some Puzzles," in Anthony Flew, ed., *Logic and Language, First Series* (Oxford: Basil Blackwell, 1951), p. 37–54).

That the controversy between the skeptic and the friend of paradigms involves the inferential/referential dialectic was clearly perceived by some of the philosophers who took part in the debate on paradigm case arguments. For example, J.W.N. Watkins pointed out that the arguments' validity depended upon giving undue priority to a term's denotation, as opposed to its "connotation": for if it is allowed

> that the term's connotation is prior to its denotation, then [the friend of paradigms]'s assertion that the term is instantiated will indeed be a factual but also an *arguable* assertion whose truth or falsity depends on how the world is constituted .... Once it is granted that the connotation of a term is primary ... the possibility will remain that the alleged paradigm examples ... were bogus examples which did not fit the specification.[16]

In other words, if being Q_1 & Q_2 &...& Q_n matters for being a *P*, then the alleged paradigms may or may not be *P*s, depending on whether they are Q_1 & Q_2 &...& Q_n; and it may even turn out that nothing is a *P*—or even that nothing *can* be a *P*, if the specified conditions are inconsistent. That the paradigms are *called* '*P*', and paradigmatically so, is, in this (inferentialist) perspective, immaterial. This view is just the reversal of what appears to be assumed by friends of paradigms,[17] and it is equally onesided (and, so far, ungrounded). However, that such a view was put forth shows that what I take to be the crucial point concerning paradigm case arguments did not go unperceived: the issue is whether we ought to be referentialists, like the friend of paradigms, or inferentialists like the skeptic, about being a *P*.[18]

---

[15] The terminology comes from my *Lexical Competence* (Cambridge: MIT, 1997). In the present context it is, I believe, self-explanatory.

[16] Watkins, "A Reply to Professor Flew's Comment," *Analysis*, XVIII (1957): 41–42. On similar grounds, Williams insisted that the argument should be limited to words that cannot be verbally defined ("More on the Argument of the Paradigm Case," p. 276).

III. PARADIGM CASE ARGUMENTS EVENTUALLY DEFEATED

We saw that validity of paradigm case arguments presupposes acceptance of the principle

(B) For every *x*, if *x* is paradigmatically called '*P*' then *x* is correctly called '*P*'.

If (B) is rejected, there is no particular reason to conclude that *o* is a *P* from the fact that people call it '*P*': people might be wrong. Now, if (B) is to be *prima facie* plausible we must be careful in characterizing paradigmatic uses. For example, it will not do to characterize paradigmatic uses simply as *regular* or *ordinary* uses.[19] Perhaps whales are regularly described as 'big fish', though such a use of 'fish' is not correct. Spiders are ordinarily called 'insects', though they should not be so called. And so forth. The intensional adjective 'so called' appears to serve the purpose of highlighting just such a gap between ordinary use and correct use: a so-called *P* is something that is ordinarily, though not correctly described as a *P*.[20]

If (B) is to have any plausibility, paradigmatic uses must be characterized in terms of semantic competence: an application of '*P*' to *o* is a paradigmatic use of '*P*' only if a speaker who is not prepared to apply '*P*' to *o* (in standard circumstances) would be regarded as semantically incompetent as far as '*P*' is concerned. Now, if a use of '*P*' fits this condition—it may be wondered—how could it fail to be a correct use of '*P*'? If *o* is an example, not just of what we *de facto* call '*P*' but of what we *ought to* call '*P*' if we are to be considered competent, how could it be incorrect to call it '*P*'? How could competent application fail to be correct application? Clearly, paradigm case arguments must be valid in principle: at most, it may be doubted that they apply to certain words or phrases. For example, one may doubt that there really are acts that deserve the title of 'paradigms of free will', intuitions in that area being notoriously shaky. But whenever it is reasonable to assume that certain uses of '*P*' *are* paradigmatic, then skepticism about the existence of *P*s is out of the question. So, perhaps paradigm case arguments were too soon dismissed, back in the 1960s.

---

[17] For example, Black, after having shown that there are paradigmatic uses of 'making something happen' and therefore of 'causing something to happen', concludes that, the skeptic notwithstanding, "it is certain ... that the notion of cause is not self-contradictory" ("Making Something Happen," p. 36): which entails that any conditions one might think of placing on being a cause are impotent against the fact that certain acts are paradigmatically called 'acts of causing'.

[18] Soames also points out that the skeptic's is what I call an inferentialist move: when she, for example, denies that there is knowledge of the external world, her point is that "the conceptually necessary conditions for knowledge are not satisfied by alleged knowledge of the external world" ("Malcolm's Paradigm Case Argument," p. 166).

[19] Malcolm, "Moore and Ordinary Language," p. 14.

[20] So the claim that "Ordinary language *is* correct language" (Malcolm, "Moore and Ordinary Language," p. 15) must be rejected, unless many qualifications are added.

However, it should be clear at this point that things are not as simple as that. Principle (B) is less unobjectionable than it appeared to be to friends of paradigms. We saw a reason why it may be thought to fail: an object, *o*, to which '*P*' paradigmatically applies may nevertheless only *appear* to be a *P*, while it is not *really* a *P* (so that it would not be correct to call it '*P*'). We argued that the objection presupposes inferentialism about being *P*: *o* is said to be only apparently a *P* for it lacks certain features that are regarded as conditions for being a *P* (for example, an elephant is an animal, not a metal-and-plastic counterfeit). On the other hand, the friend of paradigms—we noticed—never claimed that *P*s were, or were not Q_1 & Q_2 &...& Q_n: he only claimed that there were *P*s.

However, does this presentation of the friend of paradigms' view fit his philosophical motivation? Suppose the friend of paradigms is out to prove, against the skeptic, that there are genuine causal connections. Presumably, he is interested in such an argumentative effort because of the role the notion of cause plays in our intellectual framework, and more generally in our life. If 'cause' were a mere label that is attached to certain acts, processes, or relations without any particular *rationale*, so that there is no connection between the notion of cause and other notions—such as time, nature, or responsibility—then the friend of paradigms' effort would be philosophically empty. For example, what would be the point of proving that there are, or have been miracles if miracles are *not* understood to be supernatural events, but just ordinary events that happened to be labelled 'miracles'? Clearly, this is not what the friend of paradigms is up to. He wants to show that there are causes *in the customary sense of 'cause'*, where the customary sense is not characterized *just* in terms of the word's application to certain acts or processes: it includes associations of the word with other words, of the notion with other notions. For the friend of paradigms' argumentative effort to have a point, he must himself be an inferentialist to some extent: otherwise he will only succeed (at best) in proving that a label has applications, not that a concept is instantiated.

Consequently, the friend of paradigms cannot evade responsibility towards the inferential associations of 'cause'. When the skeptic grants that '*P*' is paradigmatically applied to certain objects or events but insists that, nevertheless, some or all alleged paradigms lack property Q_i which is a necessary condition for being a *P*, the friend of paradigms cannot treat the objection as in principle irrelevant; he must show either that the paradigms do have Q_i, or that Q_i is not a necessary condition for being a *P*, nor is it part of the "customary sense" of '*P*'.

It may seem that the friend of paradigms can easily meet the first requirement. When presented with a condition Q_i that the alleged *P*s must fit, he can insist that, indeed, they do fit it—and that no matter what Q_i is.[21] For example, were the skeptic to point out that the alleged paradigmatic elephant has turned out not to be an animal (it is in fact an elaborate contraption of circuits and wiring), the friend of paradigms would just retort that it *is* an animal: that big grey thing is exactly what we call an animal. Were the skeptic to insist that animals are supposed to be living organisms, not artificial devices, the friend of paradigms would reply that *that* is just what we call a living organism; and so on. What keeps the friend of paradigms from going through this routine of obstination? The answer is, his own allegiance to principle (B). According to principle (B), the ground for calling an application of '*P*' to *o* 'correct' is that it is a *paradigmatic* application of '*P*'. However, although this elephant may have been called an animal, artificial devices are *not* paradigmatically called 'animals'. Once it is found that the big grey thing in question is an artificial device, it no longer counts as a paradigmatic animal: it cannot be integral to semantic competence to call it an animal (or a living organism, and so on). Therefore, the conclusion (via principle (A)) that it is an animal is ungrounded, for it cannot be concluded (via principle (B)) that it is correct to call it an animal. The friend of paradigms cannot get rid of the skeptic by way of obstination.

Therefore, being paradigmatically called '*P*' does not unqualifiedly entail being correctly called '*P*' (principle (B) fails): there are additional constraints on the correct use of '*P*', which may be inconsistent with paradigmatic use. When the skeptic grants paradigmatic use but still claims that there are no *P*s, she is indirectly pointing out such an inconsistency.[22] Of course, in so doing the skeptic may be relying on a definition of '*P*' which is not an obvious explication of the word's "customary sense," that is, of the inferential associations that are part of ordinary competence: in that case, the skeptic will either have to show that her definition is legitimately derived from the word's customary sense, or she will have to concede to the friend of paradigms that she is using the word in a nonstandard sense. In the latter case, the friend of paradigms wins. But he never wins *just* by showing that there are paradigmatic uses of '*P*', for—to repeat—his argumentative enterprise has a point only if he is responsible towards '*P*''s obvious inferential associations; he is not entitled to take (B) for granted. Paradigm case arguments are thus either pointless or invalid. If (B) is unconditionally assumed, that is, if no responsibility is undertaken for '*P*''s inferential associations, then the argument is valid but pointless, for there is no point in showing that there are *P*s unless '*P*' is understood in the customary sense. If on the other hand (B) is not assumed then the argument is invalid, for without the help of (B) it does not follow, from the fact that *o* is paradigmatically called '*P*', that there are *P*s.

---

[21] This objection was inspired by a remark of Paolo Casalegno.

[22] Though in a misleading form, for she does not make it clear that by 'There are no *P*s' she means 'There are no things that are $Q_1\&\ldots\&Q_n$, *consequently* there are no *P*s'.

IV. PUTNAM, KRIPKE, AND REFERENTIALISM ABOUT BEING *P*

Friends of paradigms did not become extinct in the early 1960s. Remember Hilary Putnam on robotic "cats":

> Even if cats turn out to be robots remotely controlled from Mars we will still call them 'cats'.... Not only we will still *call* them 'cats', they are cats.[23]

The reason Putnam initially gave for this pronouncement was that 'cat' refers to the entities that were believed to fit the original theory associated with 'cat', even though the theory is false and the entities do not fit it (and they are no longer believed to fit it) (*ibid.*, p. 143). Later, Putnam came up with a different account of why cat-like robots are cats: according to the new, now standard account, being a cat is having the same nature—to be identified by natural science—as the *local* cats (also called 'paradigm cats'), that is, the cats around here, the objects we use to call 'cats'.[24] Of course, it may turn out that some of the alleged paradigms are anomalous in some important way: "future investigation might reverse even the most 'certain' example."[25] In that case, the anomalous examples do not count: they do not contribute to determine what it is to be a cat. So, being a cat is having the same nature as *most* of the beings we have been calling 'cat', or, equivalently, having the same nature as *this* being, *provided* it shares the nature of most of the beings we have been calling 'cats'.[26] Definitions, on the other hand, are irrelevant: our customary criteria (what we might identify as the "ordinary sense" of 'cat') do not determine what it is to be a cat; indeed, they may apply to things that are not cats and fail to apply to some genuine cats. Even the most up-to-date scientific specification of what it is to be a cat is defeasible (*ibid.*): further investigation may find that it does not really pick out cats and only cats.

---

[23] Putnam, "Is Semantics Possible?" p. 143.

[24] Putnam, "The Meaning of 'Meaning'," pp. 232, 234.

[25] Putnam, "The Meaning of 'Meaning'," p. 225.

[26] A thorough discussion of the precise form this should take can be read in Nathan Salmon, *Reference and Essence* (Princeton: University Press, 1981), chapter 4 (final formulation on p. 139).

Putnam's notion of a paradigm does not quite coincide with the friend of paradigms' notion. For the friend of paradigms, *o* counts as a paradigmatic *P* if being disposed to call *o* '*P*' is a necessary condition of semantic competence. For Putnam this is not sufficient. There probably was a time when whales were regularly called 'fish'; indeed, calling a whale a fish may have counted as part of semantic competence relative to 'fish'. Or so we may assume. Even then, no whale counted as a paradigmatic fish in Putnam's sense, for no whale shared the nature of most objects that were called 'fish'.[27] To be a paradigmatic *P* in Putnam's sense something must share the nature of most objects that are normally, or regularly, or perhaps paradigmatically (in the friend of paradigms' sense) called '*P*'.

Consequently, principle (B) does not accomodate Putnam's view if 'paradigmatically' is read in the friend of paradigms' sense:

> (B) For every *x*, if *x* is paradigmatically called '*P*' then it is correct to call *x* '*P*'.

In fact, whales were paradigmatically, though not correctly called 'fish'. On the other hand, if something is a paradigm *P* in Putnam's sense then it is a *P* and it is correct to call it '*P*'. Thus, (B*) holds:

> (B*) For every *x*, if *x* is paradigmatically* called '*P*' then it is correct to call *x* '*P*',

where something is paradigmatically* called '*P*' just in case it is called '*P*' and it shares the nature of most objects that are normally called '*P*'.

Saul Kripke's view does not differ:

> Cats might turn out to be automata, or strange demons ... planted by a magician. Suppose they turned out to be a species of demons. Then ... the inclination is to say, not that there turned out to be no cats, but that cats have turned out not to be animals as we originally supposed. The original concept of cat is: *that kind of thing*, where the kind can be identified by paradigmatic instances. It is not something picked out by any qualitative dictionary definition.[28]

---

[27] I owe this remark to Andrea Bianchi.

[28] Kripke, *Naming and Necessity* (New York: Blackwell, 1980), p. 122.

The notion of being a cat is fixed by paradigmatic instances, things we call 'cats'. Even if they turned out not to have the properties by which we identify them (including the essential property of being living beings) they would still be cats. It appears that Kripke is siding with Putnam and the friends of paradigms: being (paradigmatically*)[29] called 'cat' suffices for being a cat. The skeptic is wrong to assume that it may not be correct to call 'cats' beings that lack some crucial property, for example, the property of being a living being.

But on the other hand suppose that, counterfactually, the mountains and other areas where we actually find gold contain a substance that has all the properties that are attributed to gold but a different atomic number—80, say; suppose that same substance is also found in wedding rings, coins, and all objects we now regard as being made of gold; and suppose further that the stuff is called 'gold'. Would we describe this as a situation in which *gold* has atomic number 80? No, says Kripke:

> We would instead describe this as a situation in which a substance... which is not gold, would have been found in the very mountains which actually contain gold and would have had the very properties by which we commonly identify gold (*ibid.*, p. 124).

So it seems that being called 'gold' (paradigmatically or otherwise) is not enough for being gold: what matters is having a certain nature, supposedly singled out by atomic number. It then seems that Kripke is really taking sides with the skeptic: even granted that there are paradigmatic uses of 'gold', gold might not exist: that is, there might be a situation in which the stuff we call 'gold' is not really gold. Can these views be reconciled?

Let us recall, first of all, that the above quoted "paradigmatic" view is not properly phrased, according to Kripke himself. Given that cats *are* animals (and that 'turn out' is a factive verb), it could not *turn out* that cats are automata or demons. For 'Cats are animals', if true, is a necessary truth: there are no possible circumstances in which cats are not animals, and therefore no circumstances in which they are discovered or turn out not to be animals. That is not a possibility. What *is* a possibility is for us to be in exactly the same epistemic position in which we are with respect to beings that are cat-like demons, not animals. Similarly with gold:

> The loose and inaccurate statement that gold might have turned out to be a compound should be replaced (roughly) by the statement that it is logically possible that there should have been a compound with all the properties originally known to hold of gold (*ibid.*, pp. 142–43).

---

[29] From now on I will drop the '*', as it is understood that the relevant notion of a paradigm is Putnam's and Kripke's, not the friend of paradigms'.

Taking this correction into account,[30] Kripke's view on the relation between being called 'gold' and being gold can be stated as follows: what we *as a matter of fact* call 'gold' is gold and cannot fail to be gold (with the usual proviso ruling out anomalous examples), though we might have called 'gold' a different substance. But, isn't *having atomic number 79* crucial—essential—to being gold? Doesn't *that* count, rather than being called 'gold'? It certainly does, in the sense that, as things stand, nothing is gold unless it has atomic number 79; however, as things stand (most of) what we *de facto* call 'gold' could not fail to have atomic number 79, for we call 'gold' *this* stuff, and *this* stuff has atomic number 79 and has it necessarily.

So, eventually Kripke (like Putnam) is a referentialist about being *P*, if '*P*' is a natural kind term: if something is paradigmatically called '*P*' it cannot fail to be a *P*. However, the relevant notion of paradigmaticity (paradigmaticity*) embeds a peculiar form of inferentialism. There is a condition that candidate *P*s must meet beyond being called '*P*', namely, sharing the same nature as most objects have that are ordinarily called '*P*'. Being ordinarily called '*P*', in itself, is only a *prima facie* sufficient condition for being a *P*. But on the other hand, the relevant necessary condition is based on a semantic fact: what must be shared in order to be a *P* is having the same nature as most objects *that are ordinarily called* '*P*'.

V. DO NATURAL KIND WORDS MAKE A DIFFERENCE?

We saw that paradigm cases arguments, in spite of their initial plausibility, do not go through: for the proponent of a paradigm case argument is committed, by the very nature of his philosophical endeavor, to show that there are *P*s *in the usual sense of* '*P*'. For example, he intends to show that there are causes such that, if a person *A* causes *C*, then he may be taken to be *responsible* for *C*; also, causes such that, if *D* causes *E*, then *D precedes E*; and so forth. In other words, the friend of paradigms is committed to at least some of the inferential associations of '*P*'. If he does not do justice to such commitments his argument loses point: he may have shown that there are *P*s in *some* sense, but his conclusion will not grind any philosophical grain. In particular, the skeptic will not feel refuted, for her claim was that there were no *P*s in a sense of '*P*' which—it is argued—coincides with, or straightforwardly derives from the customary sense.

We saw that the proponents of what is now the standard view of natural kind words[31] are not quite in the same boat as the friend of paradigms, for they grant that putative *P*s must meet a further condition beside being (competently) called '*P*', namely, sharing the nature of most things that are currently called '*P*'. Is this really the *only* condition putative *P*s must meet? Could one object that, even in the case of natural kind terms, nothing is correctly called '*P*' unless it meets some further conditions corresponding to inferential associations of '*P*' (supposedly constitutive of the "usual sense" of '*P*')? If such were the case, having the same nature as most paradigmatic *P*s would not be sufficient for being correctly called '*P*': the standard theory of reference for natural kind words would fail. Moreover, assuming (A), the standard explication of *being P* in terms of having the same nature as most paradigmatic *P*s would also fail: Kripke's and Putnam's particular version of referentialism would be shown to be inadequate as an account of *being P*.

---

[30] The correction is endorsed by Putnam ("The Meaning of 'Meaning'," p. 233).

[31] I am disregarding objections that have been raised against Kripke's and (particularly) Putnam's account of natural kind terms and their semantics, including skepticism—now

The standard view we have been discussing concerns natural kind words such as 'gold' or 'cat', not words and phrases such as 'cause', 'free will', or 'past'. It can be argued that the limitation is crucial: while it makes sense to suppose that the use of 'cause' is partly governed by the word's inferential associations, no such assumption is in order for natural kind words. 'Cause' is not just a label we attach to certain acts or events; 'cat', on the other hand, *is* just the name of *these* things and their likes. Granted, even 'cat' has what one might want to call 'inferential associations', meaning that some beliefs about cats are widely shared: for example, the belief that cats are animals, that they are often kept as pets, that they meow, and so on. There is no reason to deny that sharing some such beliefs may be a condition for being judged competent in the use of the word 'cat': this is indeed the point of Putnam's notion of stereotype.[32] However, such inferential associations are standardly regarded as inert as far as the determination of reference is concerned: the word 'cat' may correctly apply to *x* even if *x* turns out not to fit any of the word's customary inferential associations (it is not an animal, it does not meow, and so on).

If so, then we should expect paradigm* case arguments—the analogue of paradigm case arguments for the new sense of 'paradigm'—to be entirely successful whenever natural kind words are involved. In other words, if *o* has the same nature as most objects that are normally

---

perhaps standard—about ordinary kinds being identifiable with scientific kinds. For a survey of such objections see T.E. Wilkerson, "Natural Kinds," *Philosophical Books*, XXXIX (1998): 225–33, and Stathis Psillos, *Scientific Realism* (New York: Routledge, 1999), chapter 12.

[32] Putnam, "The Meaning of 'Meaning'," pp. 247–49.

called '*P*' (for '*P*' a natural kind term) then *o* is correctly called '*P*', no matter whether it does or does not fit '*P*'s inferential associations; so that, by (A), *o* is a *P*.

Is this unquestionably so? Consider okapis, a species of large African ungulates (*Okapia johnstoni*). When they were first discovered, by Sir Harry H. Johnston in 1900, they were named *Equus johnstoni*. Later, they were found not to be related to horses but to giraffes (okapis are in fact the only known living relative of the giraffe). The name *Equus johnstoni* was dropped, and the species was renamed *Okapia johnstoni*, which remains its name to this day.[33] The reason the name *Equus johnstoni* was given up is, that it implies a relation to horses that okapis—those animals—do not have. The implication was not just an implicature or a harmless "connotation": it is a consequence of the very design of zoological taxonomy. Now, it may be argued, if in the case of natural kind words inferential associations were irrelevant the name *Equus johnstoni* would have been kept: naturalists would have said that *Equus johnstoni* had been found not to be a kind of horse. On the contrary, the name was dropped. This shows that the name *Equus johnstoni* was partly governed by its inferential associations. To be correctly called an *Equus johnstoni*, it was not enough to share the biological nature of most animals that were so called; in addition, it was also necessary to be biologically related to horses. '*Equus johnstoni*' was not a pure label designating *those* animals, whatever their nature and relation to horses or any other beast. Yet *Equus johnstoni*, being the name of an animal species, is a natural kind word: so it is just not true that the application of natural kind words is indifferent to their inferential associations.

Is *Equus johnstoni* a genuine natural kind term? It was intended to be, and it did designate a natural species, for a while. However—it may be argued—it differs from names such as 'tiger' or 'gold' in being *partially descriptive* in Soames's sense[34]; or more precisely, in standing to "regular" natural kind predicates in a relation that is the analogue of a partially descriptive name's relation to a regular proper name. The referent of a partially descriptive name such as 'Princeton University', says Soames, is partly determined by a descriptive property (for example, the property of being a university) that is part of the name's semantic content: *x* is the referent of 'Princeton University' if and only if it stands in the appropriate causal relation to the name *and*, in addition, it is a university. Analogously, *Equus johnstoni* names a *horse-related* species whose members have the same biological nature as paradigmatic okapis—*those* ungulates. If the ungulates in question turn out not to be related to horses, then they are not *Equus johnstoni*; as no other species satisfies the expression's application conditions, the name must be dropped. The name was dropped, not because okapis did not fit the name's inferential associations, but because they did not fit the name's *content*.

---

[33] This information derives from B. Grzimek, "Okapis," in *Encyclopedia of Mammals*, Volume 5, S.P. Parker, ed. (New York: McGraw-Hill, 1990), pp. 262–65. See http://www.ultimateungulate.com/Artiodactyla/Okapia_johnstoni.html.

[34] Soames, *Beyond Rigidity* (New York: Oxford, 2002), p. 51.

So, the claim that principle (B*) is valid for natural kinds can be salvaged by disqualifying scientific names of species from being genuine natural kind terms. If on the other hand we were to decide that scientific names of biological species are perfectly good natural kind terms, we would conclude that *Equus johnstoni* is a counterexample to the claim that the application of natural kind terms is not governed by their inferential associations; it would follow that the standard theory of reference for natural kind terms fails.

Our hypothesis has been that, in the case of natural kind terms, principle (B*) applies: that is, being a natural kind term is sufficient for (B*) to apply. One may wonder if the condition is necessary as well, that is, if (B*) applies *only* in the case of natural kind terms. One reason not to take the condition to be necessary is that, if it were, (B*) should not apply in the case of Putnam's robotic cats. Robotic cats, being artificial entities, do not form a natural kind (this is part of what is imagined to be discovered in Putnam's thought experiment): hence, 'cat' turns out not to be a natural kind term, if natural kind terms are defined as terms that designate natural kinds, as opposed to terms that are at some point *believed* to designate a natural kind. If, however, (B*) does not apply in the case of robotic cats, that is, if paradigmatic uses of 'cat' are not necessarily correct uses, we are left with no reason to persist in calling the robots 'cats' or to conclude that they are cats.

We may, of course, decide that Putnam and Kripke were simply wrong in their description of what we would (and should) say once we discover that "cats" are robots: they should have said that the correct conclusion would be that there are no cats. We had been told that to be a cat is to have the same nature as *these* things. However, something must be a natural object in order to have a nature. Consequently, the discovery that *these* things do not have a nature in the appropriate sense seems to entail that nothing is a cat. Or so one might argue.

If, on the other hand, we are persuaded by Putnam's and Kripke's description of our reactions to the discovery of robotic cats, we may give up the hypothesis that being a natural kind term is a necessary condition for (B*) to apply: as a matter of fact (we might say), the inference from paradigmatic to correct holds for other terms as well. However, we know that the inference does *not* hold in some, and probably in many cases: it would then be natural to look for some principled distinction between cases like robotic cats, in which (B*) does apply, and cases such as 'cause' or 'past event', in which it does not. The following is an attempt in that direction.

Notice that, in Putnam's thought experiment, robotic cats were originally *believed* to form a natural kind (we might say that the word 'cat' had been introduced as a natural kind term). So, it may be that this condition is both necessary and sufficient: that '*P*' is originally believed to be a natural kind term, that is, that it is believed to refer to a natural kind. The hypothesis is borne out by contrasting robotic cats with witches. Suppose there was a place, perhaps late sixteenth century Europe, where belief in witches was universal. At that time and place, being able to call certain women 'witch' was integral to semantic competence on 'witch'. Later, people became convinced that there were not, and never had been women who made a contract with the devil, conjured the dead, hindered women from conceiving, and so forth. At that point, they did not phrase their discovery as the discovery that *witches* do not make contracts with the devil, and so on; they phrased it as the discovery that there are no witches. Why? The present hypothesis provides an answer: while Putnam's cats were originally believed to form a natural kind, no such belief obtained in the case of witches. Witches were not believed to share a nature: they were (taken to be) just ordinary women that entered a particular relation with the devil. Being a witch was a matter of culture, not nature. This is why we did not persist in describing such women as witches, whereas (in Putnam's imaginary case) we would keep calling the robots 'cats'.

So far, it seems that being originally believed to designate a natural kind is both necessary and sufficient for principle (B*) to apply. Indeed, one could take this condition as a definition of 'natural kind term': a natural kind term is a term that is (originally) *regarded as* designating a natural kind, or, as Nathan Salmon says, a term that *purports to* denote a kind (*op. cit.*, p. 43). A natural kind term's semantic properties are thus made to depend, not on its actually designating a natural kind but on its purporting to designate such a kind. Such a definition would have an advantage: on the alternative definition, the semantics of a natural kind term may be forever opaque to its users, for we may never find out whether a term '*P*' is a natural kind term (for example, because we may never get rid of the mistaken belief that objects to which '*P*' is applied form a natural kind). Whereas if a natural kind term is a term that is originally *believed* to designate a kind, the linguistic community as a whole, if not the individual speaker, is supposedly aware of its semantics.

There is, however, an apparent counterexample to the present hypothesis. It seems to show that, for principle (B*) to apply, it is not sufficient for a term to be a natural kind term, not even in the sense of purporting to designate a kind. It is the example of ether. Maxwell's and Lorentz's *luminiferous ether* was taken to be an elastic, inert medium filling up universal space and carrying light.[35] Nowadays, most physicists think that ether does not exist. The term 'ether' no longer occurs in current physical theories. To the extent to which we can say that the name 'ether' was abandoned, it was because people started to believe that nothing had the properties that were associated with it: there was no inert, elastic, homogeneous, isotropic medium whose vibrations were light waves. Scientists did not say that *ether* was neither isotropic nor homogeneous, and so on, they said that ether did not exist. The name 'ether' was forsaken, *because* its defining properties were no longer believed to be instantiated. Thus—it may be argued—'(luminiferous) ether' is a counterexample to the hypothesis that principle (B*) applies to natural kind words: it is a natural kind term that was forfeited when people came to believe that its defining properties were not jointly instantiated.[36] If the counterexample is sound, even some natural kind words are governed by their inferential associations.

It has been argued that ether's defining properties were *not* shown not to be jointly instantiated: on the contrary, its "core causal description" was "taken up" by the term 'electromagnetic field'.[37] The term 'ether' may (still) be seen as referring to the electromagnetic field: there is "referential continuity" between the two terms. Why, then, was the word 'ether' abandoned? Perhaps because scientists were concerned with the word's "connotations":

> Even though the models of the elastic solid-like composition of the ether were not part of the core causal description of the term 'luminiferous ether', they can be seen as belonging to the connotations of the term. Hence, the fact that a term like 'ether' is abandoned may be explained on the ground that it was desirable that the term associated with the carrier of light-waves should avoid the connotations of 'ether'.[38]

I will not discuss the issue of how core-constitutive were the properties that were shared between the notions of ether and electromagnetic field: I am willing to grant Stathis Psillos that the two notions had important features in common. However, it is plausible to suppose that, if the term 'ether' was abandoned (as it was), its "connotations" must not have been regarded as semantically accessory. Very often in the history of scientific language (and language in general) words have been retained even though conceptions of their referents underwent deep changes: think of 'heart'. It is not plausible that, just in the case of 'ether', some accessory association was regarded as reason enough to drop the name. Clearly, some among 'ether''s uninstantiated inferential associations must have been seen as semantically crucial.

---

[35] For a short historical account of the notion (or notions) of ether, see C.A. Decaen, "Aristotle's Ether and Contemporary Science," *The Thomist*, LXVIII (2004): 375–429.

[36] Strictly, 'ether' is not a natural substance term; it is sometimes called a "physical magnitude" term, like 'electricity' (see, for example, Psillos). However, it is generally believed that such words share the semantic properties of 'gold' and 'water' (see, for example, Soames, *Beyond Rigidity*, p. 305).

[37] Psillos, *Scientific Realism*, p. 296.

[38] Psillos, *Scientific Realism*, pp. 298–99.

However, even granting that the word 'ether' was abandoned because it was concluded that nothing fitted its inferential associations, it can be argued that 'ether' is no counterexample to the hypothesis that the inference to paradigmatic to correct holds for natural kind words (that is, that principle (B*) applies). The point is simply that *there never were* paradigmatic uses of 'ether'; consequently, the conditional in (B*) is vacuously true in the case of ether, as every instance of its antecedent is false. Notice that the point is not that ether could not be perceived or pointed at. We saw (section I) that paradigm case arguments do not require that paradigms are materially exhibited: they may be linguistically described, provided the existence of something fitting the description is not at issue. The existence of ether was uncontroversial in the year 1800 (say): people who competently used the word 'ether' at that time had no doubts that they were using the word to refer to *something*—some medium—evidence for the existence of which was plentiful. But, in spite of such firm beliefs, there really was nothing there to be called 'ether', paradigmatically or not. Some entities—women—were there to be described as 'witches'; robots are imagined by Putnam to be there to be called 'cats'. But in the case of ether, *nothing* was there. People did believe that the word was used to refer to something; and, if the notion of paradigm had been around, some applications would no doubt have been regarded as paradigmatic. However, there really were no applications of the word, paradigmatic or not. There was no stuff sharing the nature of most stuff normally called 'ether', for *nothing* was ever called 'ether'. Thus, our hypothesis stands.

VI. CONCLUSION

We have been assessing the validity of principle (B*)—licensing the inference from paradigmatic use to correct use—in the case of natural kind terms. If (B*) were valid for natural kind terms, then referentialism about being *P* would be justified, limited to such terms. Kripke's and Putnam's theory of reference for natural kind terms would be vindicated. We saw that being a natural kind word is not sufficient for (B*) to be valid: the example of *Equus johnstoni* shows that a natural kind term may be dropped because the objects it is paradigmatically applied to do not fit the term's inferential associations. However, *Equus johnstoni*, being partially descriptive, is not a "pure" natural kind word. The theory may only apply to pure natural kind words.

Next, we saw that being a (pure) natural kind term does not seem to be necessary for principle (B*) to apply, for the principle is taken to apply to robotic cats, which do not form a natural kind. Moreover, if being a natural kind term is regarded as the only sufficient condition for (B*) to apply we are left without an explanation of why we would persist in describing the robots as 'cats'. Noticing that robotic cats are assumed to have been originally believed to constitute a natural kind, we put forth the hypothesis that being originally believed to be a natural kind term may be the condition, both necessary and sufficient, for the inference from paradigmatic to correct. The hypothesis was borne out by the example of witches: the name 'witch' was dropped when it was found that nothing fitted the word's inferential associations, so principle (B*) was not applied; but then, witches had never been believed to form a natural kind. The seeming counterexample of 'ether' was disqualified by showing that, if present science is right, no paradigmatic application of 'ether' ever took place.

It thus seems that the standard theory of reference for natural kind terms is vindicated, with a couple of qualifications. First, the theory only applies to *pure* natural kind and natural substance words: partially descriptive noun phrases such as *Equus johnstoni* are excluded. Second, the theory applies to words that are originally *believed* to designate a natural kind, whether or not they eventually turn out to designate one. Alternatively, we might stick to the unqualified claim that the theory applies to (pure) natural kind terms and give up Putnam's example: we would then describe the discovery that cats are robots as the discovery that there really are no cats, and there never were any (perhaps on this particular example intuitions are split anyway.) We would thus lose the contrast of robotic cats and witches: in both cases, we would be drawing the conclusion that there are no such things.[39] The "qualified" option, on which the theory applies to terms that purport to designate a natural kind, has the advantage of making the semantics transparent to intentional use: whether a word has the semantics of a natural kind term turns out to depend on whether it is *meant* to designate a natural kind, not on whether it *really* designates one. On the "unqualified" option, we may never know what semantics a word really has, for we may never know whether it really designates a natural kind.

---

[39] Notice that even in the case of witches intuitions are shaky: the statement that *witches* were found to be just ordinary women does not necessarily sound preposterous.

According to the standard view, if a word is a (pure) natural kind term its inferential associations do not contribute to determine its reference. It follows that paradigm* case arguments will go through: there is no room for skepticism about elephants or gold. Does this conclusion vindicate paradigm case arguments in general? If we look at the arguments' intended point, the answer is no: in actual philosophical practice, they were meant to refute skepticism about free will, causes, or the reality of the past, not skepticism about cats or water. But we saw that outside the limited domain of natural kind and natural substance terms, the arguments are either invalid or utterly ineffective.

Discussion has shown that paradigm case arguments gave voice to two very powerful intuitions. The first is the reduction of meaning to naming; it is the intuition which has come to be known (by way of Wittgenstein's *Philosophical Investigations*) as Augustinian. According to this intuition, a natural language word's role reduces to its naming one or more objects. The intuition is, in general, misguided: a word's semantic role is, in general, constrained by its association with other words. However, as far as pure natural kind and natural substance terms are concerned, the intuition may be correct. In this case, connection with paradigmatic* instances appears to outweigh any other associations the word may possess (perhaps even the association of 'cat' with 'animal').

The other intuition is the "referentialist" intuition on which being called '*P*' suffices for being a *P*. Again, on the face of it the intuition is wrong: being called '*P*' is not sufficient for being a *P*; being *correctly* called '*P*' is. The adverb 'correctly' hints at factors that constrain the application of '*P*' over and beyond the fact that '*P*' is ordinarily, or even paradigmatically used to refer to something. Something that happens to be called '*P*' is a *P* only if such constraints are satisfied. Nothing is a cause unless it temporally antecedes its effect, nothing is a microscope unless it magnifies very small bodies, and so on. In the case of natural kind terms, it would seem that no such constraints are active: no first-order property is a necessary condition for something being gold, or a cat. There are, however, second-order properties that represent such conditions: sharing the nature of most instances of gold is a necessary condition for being gold. Thus, even in the case of natural kinds being called '*P*' is not sufficient for being a *P*: pure referentialism fails.

DIEGO MARCONI

University of Torino

THE FOUR CAUSES*

It is usually assumed that when Aristotle distinguishes between the four causes[1] a natural scientist should study, he is referring to (1) the matter and (2) the form of a thing, (3) that which brought it into existence, and (4) its purpose. On this basis, it must remain obscure what the four items in this list have in common such that all and only the four of them should be called "causes." In what sense does the form or the purpose of a thing *cause* this thing? What is the *effect* of matter or form? And why should there be exactly four causes? For instance, why is not the weight of a thing a fifth kind of cause; and if the reason is that the weight of a thing is one of its forms, why should the material constitution of a thing not also count as one of its forms? Moreover, many properties of a thing appear to be reducible to its purely material features. Could we perhaps reduce all of them? It seems that we might do without either material or formal causes. Above all, the final cause does not appear to belong in a list of causes that all natural things have. In large areas of modern science, we do not refer any longer to purposes, and where we still do so, we try not to.

When Aristotle presents his system of four causes, he does so without arguing for it. It has therefore been claimed that we "do not know how Aristotle arrived at the doctrine of the four causes."[2] In this paper, I will point out that there is in fact a fairly simple rationale for his scheme. This rationale explains Aristotle's division; it need not reflect the way in which Aristotle arrived at it. My aim is to say what Aristotle said, not in the sense of reporting what he said, but rather in the sense of repeating and elaborating on what he said. I will show that his fourfold distinction naturally arises from the combination of two distinctions that apply to all natural beings. First, concerning any natural change, we may distinguish between the thing that changes and the

---

* Work on this paper was made possible by the Volkswagen Foundation, the Alexander von Humboldt Foundation, and the Deutsche Forschungsgemeinschaft. I have benefited from comments and suggestions by Klaus Corcilius, Allan Gotthelf, James Allen, Catherine Stinson, Stephan Schmid, Michael Thompson, and an anonymous reviewer.

[1] I use the term "cause" here only because it is the standard translation of αἰτία. I pretend not to know what "cause" means, so that the meaning may emerge from the ways in which Aristotle actually uses the term. The labels "material cause," "formal cause," and so on, are not used by Aristotle. I also use them without putting any weight on their literal meaning.

[2] William David Ross, ed., *Aristotle's Physics* (New York: Oxford, 1936), p. 37.

0022-362X/09/0603/137–60

change that it undergoes. Neither of these could be studied without in any sense referring to the other. Second, we may ask out of what a natural thing comes to be what it is, and we may ask an analogous question about a natural change. Conversely, we may ask what a thing or a change naturally comes to be. Again, the two questions, out of what something comes to be what it is, and what it comes to be, cannot be separated. I will maintain that within the cross-classification that results from combining these two distinctions, the so-called material cause occupies the slot for that out of which a thing comes to be when it goes through a natural change, and the formal cause is what it naturally comes to be. Likewise, the efficient cause is that "out of which" a natural change comes to occur; and the final cause is what the change naturally comes to be when it occurs.[3]

I. NATURAL CHANGES

We must first clarify what it is for a change to be natural. Aristotle writes that a *substance* is natural if it has in itself a principle of motion and rest (*Physics* B1 192b13–16). He also tells us that in one sense, the typical development of a natural thing may be called its nature (φύσις, 193b12–13). A change should accordingly be natural, first, when the thing that undergoes this change has, in some sense, "in itself" a principle that governs this change. Second, a natural change should constitute a development that is typical, that is, natural *for* the changing thing.

Aristotle also says, misleadingly, that natural things differ from artifacts in that they have an innate *impulse* to change (ὁρμή, 192b18–19). This seems to imply that a change is natural to the extent that the changing thing is or contains that which initiates the change. But as Aristotle clarifies in *Physics* Θ4, that a natural thing possesses an internal principle of motion and rest need not mean that it initiates motion or rest. The principle may as well merely govern the way in which things of its kind are affected by the impact of others (255b30–31). Moreover, as Wieland has pointed out, a change may very well be natural for a thing even if the relevant causal chains (in the modern sense of "cause") begin outside this thing.[4] By showing, for instance, that the movements of an animal have causal antecedents outside this animal, one cannot prove that they are not natural for it (cf. *Physics* Θ6, 259b1–20). Otherwise, we would have to assume that all natural movements are spontaneously initiated by the things which undergo them, by what Kant would have called "causality of freedom." In order to avoid this counterintuitive result, we had better not assume that the principle of motion and rest Aristotle is referring to is that which initiates motions or stops them. Rather, what he seems to have in mind is a principle that determines whether a change is natural for a thing, and whether the course it takes is typical of a change that is natural for this thing. If a change proceeds in accordance with such a principle, it may be called natural *for* the changing thing. That a principle of motion and rest is "in" a natural thing then means that the criteria according to which its movements and its development may be assessed as natural or typical arise exclusively from a proper account of its nature. The principle is not so much in the changing thing as it is in the correct account of its natural kind. We need not look beyond this account in order to see whether what happens with the thing is natural or not.

---

[3] Wolfgang Wieland, *Die aristotelische Physik* (Göttingen: Vandenhoek and Ruprecht, 1992), pp. 111 and 266, has claimed that in *Physics* B3, Aristotle only discusses the causes of persistent things. But Aristotle could not discuss the causes of natural things without at the same time considering the ways in which they change. Allan Gotthelf, in contrast, supports the position defended here by arguing that "coming to be for the sake of is ... prior to being for the sake of"—"The Place of the Good in Aristotle's Natural Teleology," *Proceedings of the Boston Area Colloquium in Ancient Philosophy*, IV (1988): 113–39, on p. 136. Against Gotthelf, David Charles has maintained that although a reduction of the final causes of things to those of processes would be possible, "it appears" that Aristotle had not attempted such a reduction and that it would constitute a "major task"—"Teleological Causation in the Physics," in Lindsay Judson, ed., *Aristotle's Physics: A Collection of Essays* (New York: Oxford, 1991), pp. 101–28, on pp. 105–06. The reduction, however, can be accomplished in a single sentence: Things and states have final causes insofar as they are, typically or as a matter of intention, involved in processes that have final causes.

As a consequence, an artifact that moves by itself does not thereby undergo a change that is natural for it by Aristotelian standards. This would only be the case if one could give a self-contained account of this artifact that implies a principle according to which its movements qualify as typical or successful. But artifacts are, by definition, designed and maintained from outside, and whether they work as they should will always depend on what their designers or users want them to do. There is nothing *intrinsically* wrong with an artifact that does not work. In this sense, artifacts behave according to external principles of motion and rest, even if they move by themselves.[5] In contrast, whether natural things "work" as they should may be decided by applying standards that are exclusively determined by their own nature. This is why we apply the concept of health to animals and plants, but not to artifacts. For to be healthy is to be in good condition according to intrinsic standards.

---

[4] Wieland, *Die aristotelische Physik*, p. 234.

[5] That something is an artifact implies that it is subject to external standards. If there were self-replicating artifacts, their offspring would cease to be artificial to the extent to which their form is not any longer determined by external standards.

That the principle of motion and rest follows from the nature of a thing means that it accounts for the way in which the behavior of this thing is characteristic of it. For instance, what a beaver does when it is building a dam is subject to a certain standard according to which it may count as typical or appropriate. This standard is tied to the specific nature of beavers. A cat would not do anything typical when building a dam, and it would not be clear what movements would be natural and appropriate for a dam building cat. Therefore, what a beaver does when it is building a dam is a natural change, whereas what a cat would do when building a dam would not be a natural change. The principle according to which beavers build their dams is not a universal and external law that would also apply to cats, but a standard that is specific and inherent to the natural kind of beavers.

I have said that a change counts as natural, first, when the account of the changing thing includes a principle of motion and rest; and second, when the change is natural *for* this thing. Since the principle of motion and rest inherent to the changing thing is that according to which the change may be said to be natural or not, both requirements amount to the same. A change is natural for a thing if and only if the account of the thing includes or implies a principle of motion and rest according to which it may be judged to be typical or successful. Since this principle determines what changes are typical for a natural thing, and since "nature" may also be taken to refer to the typical development of a thing, the principle of motion and rest inherent to a thing will also determine its nature in this sense.

The following are examples of natural changes. According to Aristotle, when a beaver comes to be, matter that formerly made up a certain amount of menstrual fluid changes into the body of a beaver. When a beaver comes to be mature, a beaver kitten changes into a mature beaver. Both are natural changes that beavers undergo. Further, when a beaver builds a dam, a pile of wood changes into a wooden dam. This is also a natural change, since it is natural for a beaver to build a dam out of wood. It is not, however, natural for the wood to turn into a dam. Dams are artifacts, and the principles of dam building apply to a beaver's dam only derivatively, by applying to the motions of a dam building beaver. In *Physics* B3, Aristotle discusses four possibly illuminating questions that a natural scientist may and should ask about such natural changes.

II. WHAT OUT OF WHICH THE THING COMES TO BE

A first question that one may ask about any natural change is out of what the resulting thing comes to be what it is (τὸ ἐξ οὗ γίγνεταί τι, 194b24). The beaver kitten comes to be out of certain stuff, the mature beaver comes to be out of a beaver kitten, and the dam comes to be out of wood. However, to determine the required meaning of the phrase "out of ..." is more difficult than it might seem. For instance, a book may be out of paper, out of letters, it might be written out of a certain desire, or it may be out of print.

In a first attempt to single out the relevant sense of "out of," we may take the "cause out of which" to be that of which the changing thing *consists*. But Aristotle also says that in the relevant sense of "out of," the conclusion of a syllogism may be said to come to be out of its premises (195a18–19). Since it is possible to derive a true conclusion from inconsistent premises, that out of which the conclusion comes to be cannot be that of which it consists. A true statement does not consist of a contradiction in any meaningful sense of "consist." Thus when Aristotle says that the "cause out of which" is "present in" the result of the change (ἐνυπάρχον, 194b24), he cannot literally mean that in all cases the resulting thing still consists of it.

Alternatively, one might want to say that that out of which a thing comes to be is that which makes its continued existence possible, such that it must persist as long as the thing persists.[6] Consider, however, the stuff out of which a beaver comes to be. Since living beings constantly exchange their matter, there will be less and less of this stuff left in the growing beaver, and once it has left the organism, the beaver in no way depends on it for its continued existence. If the matter of the beaver were that out of which it came to be, and at the same time that on which it still depends for its existence, we would have to admit that beavers gradually dematerialize during their lives. For there will be less and less of the stuff left that is both that out of which they came to be and that on whose presence they still depend.

Note, however, that Aristotle does not simply equate the "cause out of which" with the matter of a given thing. He rather lists matter (ὕλη) only as one possible instance (195a17). It is thus possible that for some natural changes, that out of which the resulting thing comes to be is not its matter. If this is so, we need not suppose that there is a parcel of matter that persists during *all* the changes in the life of a living being. We may solve the puzzle of the dematerializing beaver by distinguishing different kinds of change and focusing on one kind of change at a time. Sometimes a beaver's matter changes its form;

---

[6] This is not exactly analogous to the case of the syllogism. The premises of a syllogism are sufficient, but they need not be necessary to establish the conclusion. The suggestion above is that the matter of a thing might be what is necessary, but not sufficient for the existence of a thing. In fact, the matter of a thing is neither necessary nor sufficient for the continued existence of this thing.

CENTRAL LIBRARY UNIVERSITY OF CALCUTTA

sometimes, the beaver exchanges its matter. However, if that out of which a thing comes to be is not in any case the stuff of which it consists, we need a more general account of what it is. In *Metaphysics* H6, Aristotle indicates the lines along which such an account should proceed. That out of which a thing comes to be when it undergoes a natural change is *that which potentially is the result.* That which actually is the menstrual fluid is also potentially the beaver embryo. This embryo and the beaver kitten are potentially the mature beaver, and that which actually is a pile of wood and other materials is potentially a dam.

We need to distinguish two kinds of potential here. There are potentials that are, as it were, used up when they are realized. For instance, a couple has the potential to divorce, but a divorced couple no longer has this potential. Other potentials are not used up when realized. When a beaver actualizes its potential to build a dam, it retains this potential. Aristotle emphasizes that the cause out of which a thing comes to be is not used up but remains present in the thing.

There are two senses in which that which results from a natural change still is that out of which it came to be. First, an actual *F* is also a potential *F.* The result of a natural change is thus still the same as the thing out which it came to be. Both are potentially the result; one of them is it *merely* potentially, the other one is potentially the result because it is actually the result.[7] The mature beaver is still the same as the beaver kitten, insofar as they both are potentially a mature beaver. The difference is only that the beaver kitten changed from merely potentially being something to potentially *and* actually being that thing.

Second, in *Metaphysics* H6, Aristotle writes that "the proximate matter and the form are one and the same thing, one potentially and the other actually" (1045b17–19). Sally Haslanger considers three interpretations of this passage: matter and form may be "one and the same" by being (1) identical, (2) two aspects of the same, or (3) two parts of the same.[8] She argues convincingly against (1) and (2). This, however, does not imply that (3) is correct. Rather, Aristotle's statement in H6 has the logical structure of, for instance, "George Bush and Barack Obama are the same: one formerly, the other presently" (uttered in 2009). It does not follow from this that Bush and Obama are identical, two aspects of the same or two parts of the same. In the same sense, *A* may be the same as *B*, one potentially and the other actually; and still, *A* and *B* need not be identical, or two aspects or parts of one thing. When a potential is actualized, it is not *combined with* its actualization; and when matter is formed, it is not combined with its form. It actually becomes what it potentially was. It follows in any case that the matter and the form of natural things are not two numerically different things. The matter of a natural thing is not one of its components, parts, or constituents.

---

[7] Cf. Mary Louise Gill, *Aristotle on Substance: The Paradox of Unity* (Princeton: University Press, 1989), pp. 178–80. In his commentary on *Metaphysics Z and H* (New York: Oxford, 1994), David Bostock reads "merely potential" where Aristotle says "potential" (p. 225), rendering Aristotle's doctrine nonsensical (pp. 283–84).

[8] Haslanger, "Parts, Compounds, and Substantial Unity," in T. Scaltsas, D. Charles, and M. L. Gill, eds., *Unity, Identity, and Explanation in Aristotle's Metaphysics* (New York: Oxford, 2001), pp. 129–70.

III. WHAT THE THING COMES TO BE

The matter that makes up a certain portion of menstrual fluid comes to make up a beaver. The beaver comes to be a mature beaver. A pile of wood and other materials comes to be a dam. Taking, in each of these cases, that as given out of which the respective thing comes to be, we may ask *what* it naturally comes to be. That out of which a thing naturally comes to be is something that potentially is this thing. What it comes to be is what it potentially is. For instance, the beaver kitten is what potentially is a mature beaver, and a mature beaver is what the beaver kitten potentially is. There are, of course, innumerable things that a portion of menstrual fluid, a beaver kitten, or a pile of wood might possibly come to be. When we ask what they *naturally* come to be, we are only interested in a certain range of these. More specifically, we are interested in what a thing would come to be as a result of a change that proceeds according to a principle of motion and rest inherent in this (or another) thing.

That into which a thing changes when it develops according to its nature has also been called its "formal cause," presumably because the specification of that into which a parcel of matter changes involves a specification of the form it takes on. This label, however, is misleading insofar as it makes us think that that into which a thing changes is a mere form. That into which a beaver kitten changes when it undergoes a natural change is a mature beaver, and a mature beaver is certainly not a mere form. That toward which a natural change takes place is rather, as Aristotle puts it in *Metaphysics* Λ3, a concrete substance or τόδε τι (1070a11).[9]

There is a sense of "form" that still applies in such cases. The *paradigmatic* form of a thing, also called its *essence*, is *what* this thing is. It is what this thing comes to be (or should come to be) as a result of its natural development. When we call a living being a "form of life," we use "form" in this sense. What a thing is, its paradigmatic form, is what its correct definition defines. The correct definition of a beaver does not define a mere form. It defines a formed compound, something that has four feet and sharp teeth. No mere form has four feet or sharp teeth.

---

[9] Aristotle says in 1070a2 that form (εἶδος) is *that into which* a thing changes; he cannot mean a mere form here. Also, when Aristotle says in *Metaphysics* H6 1045b17–20 that matter is potentially what form (μορφή) is actually, "form" cannot be mere form. Bostock diagnoses nonsense, since it follows that the form actually has the features of the formed thing (*Metaphysics Z and H*, p. 284; cf. note 7 above). But this is precisely what we must assume. (We must *not* assume that the form of a thing is a separate thing on a par with its particular instances.)

Aristotle makes it clear enough that in *Physics* B3, he is talking about the paradigmatic form (παράδειγμα) and the essence (τὸ τί ἦν εἶναι) of a natural thing (194b26–27). Since the paradigmatic form is what a thing turns into when it undergoes a natural change, the essence of a naturally changing thing is literally "what it was to be": a beaver, a mature beaver, a dam.[10]

That the paradigmatic form of a natural thing should be that which results from its natural development will strike many as extremely odd. For instance, if the paradigmatic form of a thing is what this thing is, the paradigmatic form of a compound thing must also be compound. But does Aristotle not, in *Metaphysics* Z4 and Z10, deny that the paradigmatic form of a thing involves its matter?

In Z10, Aristotle seems to imply that the paradigmatic form of a natural substance cannot be a composite of form and matter. He says that man, taken in general, is a composite of a sort *and not* an οὐσία (1035b28–30), and arguably, the paradigmatic form of a thing is what Aristotle calls an οὐσία here. What he says might thus indicate that in general, no composite can be a paradigmatic form. But nothing requires this reading. The reason why man, taken generally, is not an οὐσία may as well be that no *universal* is an οὐσία, as Aristotle proves in Z13.

In Z4, Aristotle argues that a formula by which a thing is said to be something else cannot be a proper definition (1030a10–11). For instance, a compound expression such as "white man" is not a proper definition (and does not designate an essence), since to be white is not in all cases to be a man and hence, in the formula "white man" something ("white") is said of something else ("man"). In contrast, to be two-footed is in any case to be a certain kind of animal, and to say that some animals of this kind are two-footed is not to say something *else* of them; they are not two-footed independently of being this kind of animal. For the same reason, it seems, the definition of a substance should not involve a reference to its matter, since the form is spoken of as something different from the matter (cf. Z3, 1029a23–24). But if the paradigmatic form of a thing does not involve its matter, it seems that it cannot be a compound.

---

[10] Cf. Monte R. Johnson, *Aristotle on Teleology* (New York: Oxford, 2005), p. 48. "What it was to be" would of course not be a correct *literal* translation of τί ἦν εἶναι, which is more accurately rendered as "what it is for a thing to be this kind of thing." In any case, the "was" does not allude at a past state, but rather at a possible discrepancy between what a thing actually is and what it is supposed to be.

Now it is true that when we say what a thing is, we do not refer to the particular portion of matter of which it is made. But this does not mean that we refer to a bundle of properties and features instead. For instance, when we say that Socrates is the particular human being consisting of *this* particular amount of flesh and bones, we say *who* he is, and not merely *what* he is. We say *what* he is only when we say that he is *a* human being. Still, *a* human being is not a bundle of features and properties. .

Paradigmatic forms may indeed be shared by many items: What a thing is may be what other things are as well. What this is, a beaver, may also be what that other living being is. However, we must not conclude from this that what a thing is is a shared property. Although all beavers may also share certain properties, what each of them is—a beaver—is not a property. The difference between a paradigmatic form and its instances is not the one between a set of properties and their bearer. It is, rather, the difference between a *generic* instance (*an S*) and a *particular* one (*this S*).[11] Each beaver is *an* instance of its kind, although only one of them is *this* particular instance of this kind.

The answer to the question raised above is that in one sense, a definition should not refer to the matter of the defined thing: it should not involve a reference to *this* or *that* particular matter. However, the definition of a thing may very well involve facts about the general material constitution that all things of the respective kind share (which Gill calls "functional matter").[12] We should recall in this context that natural things are what they are in virtue of possessing an inherent principle of motion and rest. Any account of what a natural thing is must refer to the ways in which such things naturally change. But, Aristotle argues in *Metaphysics* Z11, one cannot treat a thing as something that changes without considering both its material and formal aspects; that is more generally, without distinguishing between that which may come to be something and what it naturally comes to be. Hence, the contrast between the material and the formal cause is essential to all natural things (1036b28–32).[13]

---

[11] One should think that in general, "*X* is *an F*" implies "There is an *F* such that *X* is *this F*." But the "this" in the latter phrase may be taken to be either anaphoric or demonstrative. As an anaphor it would not refer to the *X* but to its paradigmatic form *F*: there is a paradigmatic form such that *this* form is what *X* is. On the other hand, if "this" is a demonstrative referring not to the paradigmatic form but to the particular instance, "an *F*" cannot be taken to refer to the paradigmatic form itself. There is still a difference between being an *F* and being this *F*.

For more on generic individuals, see Gregory N. Carlson and Francis J. Pelletier, *The Generic Book* (Chicago: University Press, 1995), and Michael Thompson, *Life and Action* (Cambridge: Harvard, 2008), chapter 4.

[12] Gill, *Aristotle on Substance*, p. 133.

Note further that the matter of a substance, considered *as matter*, is nothing but this substance by way of privation (H1, 1042b3). We can only refer to it, as matter, by specifying what it potentially is. Now as Aristotle says in Z7, the form of a privation is the form of what is missing (1032b3–4). Hence, the matter of a thing has, as such, no form of its own (cf. Z10, 1036a8–9). Its form is the form of what it potentially is, and this is (or is part of) the paradigmatic form of the substance. Although a definition may refer to the constituent parts and the functional matter of a thing, it must therefore refer to them by way of their form; but since this is the form of what they potentially are, they do not directly figure in the definition *as matter* (cf. 1035a7–9).

I conclude that the formal cause of a natural thing is not its mere form, but a paradigmatic formed compound. It is not a property or feature of the thing, and *a fortiori* not a *necessary* property or feature. Essences are not necessary properties. But while we must be very careful not to confuse the paradigmatic form of a natural thing with properties, features, or bundles of them, we must also not confuse it with a particular individual. The paradigmatic form is that which is defined, and no particular thing can be defined (as particular). Further, even a thoroughly typical particular instance of a kind would not be the same as the paradigmatic form with a view to which it qualifies as thoroughly typical. No particular instance of a kind can *be* that with a view to which its own perfection is measured. Nonetheless, whereas the paradigmatic form is not a particular instance of its kind, it is still a generic *instance* of it.[14]

---

[13] See Michael Frede and Günther Patzig, *Aristoteles 'Metaphysik Z'* (Munich: Beck, 1988). The general point that all natural things are material follows if one reads αισθητὸν in 1036b28; cf. Ernst Tugendhat, TI KATA TINOΣ (Freiburg: Alber, 1958), p. 111. If Frede and Patzig are right and one should read αισθητικὸν (*Kommentar*, p. 210), the point is that, more specifically, beings capable of sensation must consist of a certain kind of matter. In any case, that natural or sensing beings necessarily consist of matter does not imply that matter must be part of their definition (*Kommentar*, p. 212). What belongs to the *object* of a definition need not have to be mentioned in the definition itself.

[14] Frede and Patzig point out that the essence of a thing is not a feature or property (*Einleitung*, p. 44), and they claim that for Aristotle, the form of a natural thing is as particular as this thing itself (p. 48). Paradigmatic forms, however, are neither as universal as features or properties, nor as particular as concrete, spatiotemporal things. They are generic instances.

Recall once again that a change is natural insofar as it is governed by a principle of motion and rest that is implied by a proper account of the nature of the changing thing. We may now restate this as follows: Whether a change is natural for a thing depends on what the thing is, that is, on its paradigmatic form. The paradigmatic form of a thing determines the typical course of the changes that are natural for it. For instance, the change by which the beaver kitten turns into a mature beaver is natural for the beaver kitten by virtue of the paradigmatic form of beaver kittens. The beaver kitten "was to be" a mature beaver. Although it may not yet fully exemplify the features associated with this form, since it is not yet mature, "mature beaver" is its paradigmatic form, since for a beaver kitten it is natural to grow into a mature beaver. It need not be the case that a natural thing actually comes to be "what it was to be." A beaver kitten may fail to develop into a mature beaver. If this happens, it falls short of its own paradigmatic form. This shows again that the paradigmatic form of a thing cannot be a set of features it has: Nothing can fall short of its own features.

The paradigmatic form of a natural thing is what it is supposed to turn into as a result of its natural development. This means that the paradigmatic form of a thing does not change (*Metaphysics* Z8, 1033b5–8). What the changing beaver is supposed to be is clear all along, and *it* does not undergo any change. In this respect, the paradigmatic form is also not the same as the particular thing that actually undergoes the change. The paradigmatic form of a thing is what it comes to be in a natural change, but it is not that which changes and comes to be something. The paradigmatic form of a thing is specified by the definition of that thing, and the definition of a thing does not change when the thing changes.[15]

Hence, what a thing is, its paradigmatic form, differs from the particular thing itself in at least three respects. First, other things of the same type may share it, so that it is what other things are as well. Second, since the paradigmatic form of a naturally changing thing is what it is supposed to change into, things may fall short of their own paradigmatic form. The paradigmatic form sets a standard that the thing might fail to meet. Third, the paradigmatic form of a thing does not change when the thing changes. But in one sense the paradigmatic form is still what the thing is: it is not *this* particular, but still *a* generic instance of its kind.

---

[15] We may refer to the changing thing (1) as a changing compound, (2) as that which has the potential of being the result of the change and remains present in it, or (3) as what it "was to be" (cf. Tugendhat, ΤΙ ΚΑΤΑ ΤΙΝΟΣ, p. 74). Insofar as a compound is first in one and then in another state, it changes. Insofar as the same thing is potentially the result of a change, it does not change; since when the change has taken place, it still potentially is the result of the change. Finally, the paradigmatic form of a changing thing does not change, since its definition does not change. Hence, the matter and form that are involved in a natural change do not themselves change.

IV. WHENCE THE CHANGE COMES TO OCCUR

Turning from naturally changing things to the changes they undergo, we may again ask *out of* what or *whence* they come to occur (ὅθεν, 194b29).[16] That out of which a change comes to occur seems to differ from that out of which a thing comes to be, because changes occur in time. I have argued that when a thing comes to be what it potentially is, it remains the same in the sense that it still is potentially what it actually is. Like the beaver kitten, the mature beaver still potentially is a mature beaver. In contrast, the potential for a certain change to occur seems to be a potential that is used up when it is realized. When a change comes to occur, it goes through a number of subsequent stages, none of which need last throughout the entire change. Therefore, that out of which a change comes to occur does not seem to persist in the change. For instance, when a beaver is felling a tree, it is biting chips off the trunk until the tree falls, and we may say that biting chips off the trunk is potentially the same as felling the tree. It is that out of which the latter comes to be. However, when all chips are bitten off the trunk, this stage of felling the tree is over, and that which potentially was the felling of the tree lies in the past. For this reason, that out of which a natural change comes to occur may seem to be only the "first source" that initiated it (ἡ ἀρχὴ ... ἡ πρώτη, 194b29–30).

But as we have already seen, when Aristotle speaks of the principle of motion and rest inherent in a natural thing, he is actually referring to that which determines what changes are natural for that kind of thing. Since he uses the same term, ἀρχή, both for this principle and the "cause whence," that out of which a change comes to occur may also be taken to be the principle that determines which changes are natural for the changing thing, rather than the impulse that initiates them. This principle persists throughout the change.

Consider further what a beaver is doing when it is building a dam. Obviously, it does not just kick off a process that would then run on its own. Rather, the beaver continuously keeps doing things that contribute to building the dam, and it keeps adjusting the movements that, for instance, the wood makes on its own. It does not simply initiate the movements that take place, but it directs them and thus causes them to be what they are. That out of which a change comes to be what it is is not only its beginning. That which determines the natural course of the change is what the beaver keeps doing, in accordance with its nature, in order to make the change what it is.

---

[16] Cf. *De Generatione Animalium* A18, 724a21–31: in some sense, the efficient cause is "that out of which."

Aristotle also identifies the "cause whence" of a change with the *agent* that brings the change about, or even with the agent's capacity of doing so.[17] This is really the same as above. For we cannot, in this context, separate an action from its agent and the agent's capacities. What the beaver keeps doing determines the distinctive course of building a dam only insofar as it is in turn determined by the principle of motion and rest inherent in the beaver. Hence, any specification of the steps that are naturally involved in building a beaver's dam must refer to the nature of the beaver. The same motions, when performed by an animal of a different kind, would not be governed by the same principle of motion and rest. Conversely, we cannot separate the agent from what it does. We may say that an agent causes an action; but of course, it is not the mere presence of the agent that causes the action, but precisely the agent's initiative or activity that does so. A natural thing may be said to be that out of which a change comes to be only insofar as it acts (cf. *Physics* B3, 195b3–6). A beaver that does not change would not determine the course of any change.

What is commonly called the "efficient cause" of a natural change is thus at the same time (1) the agent that directs this change according to its nature and (2) what this agent does in accordance with its nature in order to do so. The agent is the efficient cause of the change only insofar as it acts, and the actions are the efficient cause of the change only insofar as they are governed by the agent's innate principle of motion and rest. The same actions done by something else need not be the efficient cause of any natural change. In the end, the efficient cause is a changing thing that causes a changing thing, or what is the same: a change in a thing that causes a change in a thing.

Note further that by doing all the things that need to be done in order to build the dam, the beaver *is* already building the dam. We may point at the beaver while it is biting chips off a tree and say: "this beaver is right now building a dam." The beaver does not perform the action of building a dam *in addition* to performing all the steps it involves. Hence in some sense, that out of which building the dam comes to be what it is is also the very same as building the dam. I have suggested, in the beginning of this section, that that out of which a change comes to be does not last throughout the change. This turns out to be partly misleading. For any step that is involved in a change, there is a description under which it does indeed last throughout the change. For instance, insofar as biting into a tree already *is* the same as felling it, it lasts as long as felling the tree does. This result is similar to the one reached with regard to the changing thing. That out of which the beaver comes to be is that which potentially is the beaver, and that out of which building a dam comes to occur is that which potentially is the building of a dam. When the beaver has come to be, the same thing that potentially was the beaver is both actually and potentially the beaver. When building the dam comes to occur, the same change that potentially is the building of a dam is also an actual dam building. This analogy is encouraging, since it might indicate that the so-called "final cause" is what a change naturally comes to be when it occurs, just as the formal cause is what a thing comes to be when it develops according to its nature. Let me therefore turn to the final cause.

---

[17] *Physics* B3, 194b30–32. In *Metaphysics* Λ3, Aristotle speaks of the moving cause (1070a21–2).

V. WHAT THE CHANGE COMES TO BE

Again, we should be able to take that out of which a natural change comes to be as given and ask *what* the change comes to be when it occurs. When a beaver is felling a tree and arranging its parts in a certain way, what the beaver is doing may be said to "turn out to be" or "turn into" the building of a dam. We may also say, on this basis, that the beaver acts *in order to* build a dam. What a change comes to be, or that into which a change turns, may therefore in a first approximation be taken to be its end, its direction, the resulting change itself, or that for the sake of which it occurs.

I have argued that the efficient cause of a natural change is not only an initial impulse. For the same reasons, that into which a change turns should not be taken to be only its final stage.[18] For instance, if a beaver leaves out one of the initial steps that are involved in building a dam, what it does might not turn out to be proper dam building. The description of what a change turns into will thus involve an outline of the entire pathway by which its final stage is reached.

I have also argued that the formal cause of a thing, if it is that into which it is supposed to change, must be a compound. We see now that in a similar sense, what a change turns into should be taken to consist of several steps and stages. If the final cause of a natural change is indeed what it comes to be, it will not only include its final stage but a sequence of steps that are typically involved in changes of this kind. When Aristotle says that it is ridiculous to identify the final cause of a human's life with her death, this may well be what he has in mind (*Physics* B2, 194a30–33).[19] For although death does indeed belong to life as its final stage, a complete life will typically involve much more than death. A human life naturally turns into a whole series of events, and not only into its end.

---

[18] If the final cause of a change is not its final stage, the eternal revolution of the stars may also be said to have a final cause. Although this movement does not end, it still has a determinate shape and typical course.

For similar reasons, that into which a change turns cannot only be its direction. The direction of a change is that into which it proceeds, but it does not determine its entire natural course. In *De Anima* B4, for instance, Aristotle refers to a change that has a direction but no final cause. The movement that fire makes is directed, primarily upward but also in all other directions, and it seems to be typical for fire to move in these ways. Nonetheless, the spread of fire has no final cause. This has nothing to do with whether the spread of fire has a *purpose* or not. What fire lacks in the first place is a limit (πέρας, 416a15–17). Since the Greek equivalent for "to proceed" (περαίνειν) derives from the same root as the word for limit (πέρας), the limit of a change may be taken to be the pattern according to which it proceeds.[20] We have already seen that in some sense, that according to which a natural change proceeds is the principle of motion and rest that arises from the definition of the thing that undergoes the change. The possibility that we are presently considering is that "for the sake of" really means as much as "according to." A change may thus be said to occur for the sake of the limit according to which it proceeds.[21] This is in fact what Aristotle says in *De Motu Animalium* 6 (700b13–6): The movements of living beings have a limit (πέρας), and this limit *is* that for the sake of which (οὗ ἕνεκα) they occur.

In *Metaphysics* Θ6, Aristotle claims that all actions that have a limit (πέρας) are only means to an end, and never themselves ends

---

[19] Aristotle writes that wine does not naturally turn into vinegar, and that humans do not naturally turn into corpses (H5, 1044b34–45a6). Obviously, he cannot mean that it is not natural for humans to die. The reason why it is not natural for them to turn into corpses is that humans and their remains do not have a common nature and that therefore, a change from one to the other cannot be governed by one and the same principle of motion and rest. It cannot be natural *for humans* to end up being corpses because there is no time at which a human being is a corpse. The human being ceases to exist when the corpse comes to be.

[20] Cf. *De Partibus Animalium* 641b23–25; cf. James G. Lennox, *Aristotle: On the Parts of Animals* (New York: Oxford, 2001), p. 146. In the *Philebus*, Socrates also establishes a close connection between τέλος and πέρας; cf. for instance 24B.

[21] Ἕνεκα may well mean "with a view to." Cf., among many possible examples, Plato, *Gorgias* 493e (ἕνεκα τούτων = so far as this is concerned, on that score); *Politeia* 548d (ἕνεκα φιλονικίας = in point of conscientiousness); 582d (ἐμπειρίας ἕνεκα = regarding experience). That a process occurs ἕνεκα *X* may thus mean that it is to be assessed as typical or complete with a view to *X*.

(1048b18–19). On the face of it, this seems to mean that the end of an action is not its limit, and that to occur for the sake of an end is not the same as to occur according to a limit. But that an action is not an end does not imply that it has none; and the end it has may very well be the same as its limit. If the end is the same as the limit, then Aristotle is claiming that actions that *have* a limit *are* not limits (at least not their own). Actions with an end do not coincide with their end, since as long as they occur, the end has not been reached. By the same token, they do not coincide with their limit, since as long as they occur, the pattern according to which they occur has not been completed.[22]

In *De Generatione Animalium* A1, Aristotle also suggests that the final cause of a natural change is its typical course as it is determined by the relevant principle of motion and rest. He tells us that "nature flees the limitless (ἄπειρον), for the limitless is without end (ἀτελές), and nature seeks an end (τέλος)" (715b15–16). The end that nature seeks when it flees the limitless should be the limit, and there is no reason for identifying this limit with a *purpose*. That nature seeks an end means that all natural changes have a limit, which is determined by the principle of motion and rest inherent to the changing thing.

That to be for the sake of something is the same as having a limit or determinate shape is also implied when Aristotle says in *Physics* B2 that all changes that *might* be due to thought are for the sake of something (196b21–22). The reason for this is that one could not possibly intend to do something entirely unspecific. We can only intend to perform an action when we know in advance what kind of action it will be (where a failed instance of a kind still counts as an instance of this kind). Hence, the things that might be done intentionally must at least have some typical and determinate shape according to which they are supposed to proceed when performed. If the final cause of a change is that according to which it proceeds, all and only such changes could possibly be intended that have a final cause. This however does not imply that a change is for the sake of something only when it is intended with a further end or purpose in view. Rather, we may only purposely intend to realize changes which already *have* a final cause.

I conclude that this is indeed how we should understand Aristotle. The final cause of a natural change is primarily the "limit" or standard pattern according to which it proceeds.[23] This limit is the (ideal) course that changes of this kind typically take, and with a view to which they are called natural or unnatural. A change that could take any turn whatsoever without ceasing to be a typical and complete instance of its kind would not be an instance of any specific kind, and it would have no determinate shape. As Aristotle says, nature would "flee" such a change; that is, no such change could possibly be a natural one. And although only changes that proceed according to a certain principle may be brought about with a purpose in view, this does not mean that the final cause according to which they proceed is a purpose.

---

[22] In Θ6, Aristotle distinguishes κινήσεις, which are complete only when they are over, from ἐνέργειαι, which may be complete before they are over. Ἐνέργειαι coincide with their own end: they may be both perfect, in the sense of complete, and ongoing. Ἐνέργειαι, however, are not natural changes. If they have a typical form, this form is more like the formal cause of a *thing* than the final cause of a change: it is the actualization of a potential that is still present in the result.

One might object at this point that Aristotle often identifies the final cause with a *good* or even the *best*.[24] Does this not imply that after all, the final cause is a good that is achieved by realizing a change? It does not; at least not in the sense that the good would be a further, external purpose. Rather, that the limit of a change is its final cause in the sense of a good or the best only means that it is the *best-case scenario* for that kind of change.[25] When the beaver bites into a tree in order to fell it, the best-case scenario is that the tree eventually falls. In this sense, that the tree falls is the best of biting into its trunk. Recall that for a beaver in that situation, biting into a tree *is* felling it. Hence, if felling the tree is indeed the "purpose" of what the beaver is doing, we must admit that this "purpose" also *is* what the beaver is doing. Hence in fact, the good that is achieved when a beaver bites into the trunk of a tree does not differ from what it does. It is not an external purpose. Accordingly, we may also say that felling the tree is the limit of biting into its trunk, and in a further step, we may say that this limit is the purpose for the sake of which the beaver bites into the tree. However, we may do so only because in this case, the "purpose" of what the beaver is doing is felling the tree, which is also what it is doing.

There is a simple reason why the limit or typical course of a natural change seems to differ from its final cause. Since we commonly already refer to changes in terms of their typical form, it will rarely be informative to state explicitly their immediate final cause. We would say, for instance, that the final cause of biting into a tree is the typical course that biting into a tree takes, and although this would be a true statement, it would not be worth our attention. It is certainly much more informative to be told that the final cause for the sake of which a beaver bites into the trunk of a tree is the typical course that *felling a tree* takes. But nonetheless, since for a beaver in that situation, biting into the trunk of a tree *is* felling it, it remains true that the final cause of what it does is the course it typically takes. To say that the beaver does something in order to do something else is only to redescribe what it does.[26]

---

[23] I do not claim that Aristotle always speaks of the final cause in this sense. I maintain, however, that the ways in which he uses the term are explicable in terms of this basic use. For instance, a process may have an *external* final cause, such that it happens "for the sake of" the "limit" of a numerically different process to which it typically contributes.

[24] See, for instance, *Physics* B3, 195a23–24: τὸ τέλος καὶ τἀγαθὸν; 194a33: βέλτιστον.

[25] In the *Nicomachean Ethics* B6, 1106b29–31, Aristotle reminds us that the bad is a form of the unlimited (ἄπειρον), and the good of the limited (πεπεράσμενον).

We may also meaningfully say that a natural thing changes for the sake of its own nature. For we have already seen that the typical course of a natural change is determined by the principle of motion and rest inherent in the changing thing. Hence, when a natural change proceeds according to its final cause, the changing thing will also behave according to its paradigmatic form. This is what it means for both to be "good."[27] Since a natural change is a change that is natural *for* the thing that changes, we may thus also say that it is "best," in the sense of *natural*, for the changing thing to change in this way and not in another. In this sense, the final cause of a change is also determined by what is best for the changing thing. What is best for a thing is what it naturally develops into.[28]

What is natural for a thing need not necessarily be pleasant for it. In some cases, it may well be "best" for a living being to suffer, namely in such cases where it is natural for it to suffer. In any case, we could not even determine what is beneficial for a given thing without already knowing what it is to be a "good" instance of its kind.[29] What is good for a beaver may not be good for a cat, since what it is to be a good beaver is not what it is to be a good cat.

So far I have argued that the final cause of a change is the limit according to which it proceeds. This limit is that with a view to which a particular change counts as a typical or atypical instance of its kind. Further, that with a view to which a change is judged to be typical is the course it would typically take. Since the final cause is the typical course that we specify when we say what kind of change it is, it is also the *essence* of the change itself, in the same sense in which the paradigmatic form is the essence of a natural thing. I will now argue further that, just as the paradigmatic form of a natural thing is a generic instance of its kind, the essence of a natural change is also a generic instance of its kind. The reason for this is that in general, there is no real difference between a change and the course it takes. When we observe a change taking its course we do not observe anything further beyond the change itself. That a change takes its course is not a further change that the first change undergoes. Hence, the typical course of a certain kind of change *is* a typical change of this kind. As a consequence, the present analysis of the final cause matches that of the formal cause given above. The paradigmatic form of a thing is what its definition defines, and although the object of the definition is not a mere form but a formed compound, it is not a particular instance of the defined kind. It is a generic and prototypical instance. Likewise, the final cause of a natural change is "what it was to be": not *this*, but *a* change of its kind. It is that which is characterized in its correct definition, and hence a generic and prototypical instance of this kind of change.[30] Moreover, I have argued that a thing may fall short of its paradigmatic form. In the same sense, a natural change may fall short of its final cause. This happens when it is interrupted before it is complete, or when it is completed in an atypical way.

---

[26] The possibility of taking the aim of an intentional action to be that very action in a broader description has been pointed out by G.E.M. Anscombe (*Intention* (Oxford: Blackwell, 1957), §26), and Thompson (*Life and Action*, pt. II). I apply their insight to natural changes in general.

[27] As Wieland puts it, "βέλτιστον ... means the inner perfection of a thing ... . ἀγαθόν or βέλτιστον is thus only a predicate that essentially applies to every τέλος as such" (*Die aristotelische Physik*, p. 264). I have already argued that the τέλος of a change is its limit.

[28] I should emphasize that I do not suppose that we may know what is natural for a thing *before* knowing what is best for it. Rather, knowing one is knowing the other and vice versa.

[29] Cf. Gotthelf, "The Place of the Good in Aristotle's Natural Teleology," p. 117. In *Physics* B7, Aristotle explains that he uses "better" *relative to a certain life form*: διότι βέλτιον οὕτως, οὐχ ἁπλῶς, ἀλλὰ τὸ πρὸς τὴν ἑκάστου οὐσίαν (198b8–9).

This is why Aristotle can say, in *De Generatione Animalium* A1, that the formal and the final cause of an item are in several respects the same (715a8–9). The reason is, first, that the final cause is for the change what the paradigmatic form is for the changing thing. Both the formal and the final cause are accounts (λόγοι) of the entity in question.[31]

Second, since the typical course of a natural change is determined by the paradigmatic form of the changing thing, the final cause of the change is also determined by the formal cause of the thing. Conversely, since the ways in which natural beings change belong to their nature, the natural changes they undergo determine their paradigmatic form. The distinction between the form of a thing and the form of its natural development is to a certain extent an abstract one. It may often seem that there is exactly one mature form of a living being, and that this is the paradigmatic form. But for a natural thing to be natural is to develop and go through several stages, such that all of these stages belong to what it is to be that kind of thing. That is, a proper account of the paradigmatic form of a natural thing will include an account of its natural development.

---

[30] I have argued that the efficient cause is not only an agent or an action, but an acting agent. For the same reasons, the final cause must include an agent. The paradigmatic instance of a kind of natural change must involve a paradigmatic instance of a changing thing. It must be a generic change of a generic thing.

[31] Cf. also *De Anima* B4 (415b14–15) and *De Partibus Animalium* A1 (639b15–16).

Despite these similarities between final and formal causes, it should be noted that they also differ in at least one important respect. There might be a thoroughly perfect and presently existing natural thing. It is, for instance, at least possible to imagine a thoroughly typical and complete beaver. However, it is for conceptual reasons impossible to imagine a perfect and still presently ongoing change; for instance, a perfect felling of a tree. For to imagine a process is to imagine it as ongoing, and as long as the felling of the tree is going on, it is not yet over. But at least for such processes as felling a tree, to be complete would involve being over, so that this process cannot be complete before it is over. On the other hand, if it is over, it is not any longer going on. Moreover, for every stage of felling a tree, at least one other stage will not be going on at present. Therefore, it is impossible to imagine a perfect and yet presently ongoing felling of a tree. In this sense, the final cause of a natural change is *external* to any concrete instance of its type. The typical course of a kind of change necessarily extends beyond any of its particular instances. It includes what has happened before each present instant and what is supposed to happen after it. For instance, when the beaver bites into the tree in order to fell it, felling the tree goes beyond biting into its trunk. This is closely related to what I have said about the efficient cause: that it does not any longer seem to be present in the change for the sake of which it occurs.

Looking back at the efficient cause, we may now see that it is the same as the final cause in two respects. First, both the efficient and the final cause are in some sense that which determines the typical course that a natural change takes. The efficient cause is that which causes the natural change to be what it is, and the final cause determines what this change "was to be." Second, when the steps occur that are involved in bringing about a change, they *are* this change. Taking steps in order to fell a tree is felling a tree. In this sense, the efficient cause is the same as the change into which it naturally turns. On the other hand, the final cause of a natural change is also what this change is, since it is its essence. It is the typical course that changes of its kind take, and as I have argued above, this is the same as a generic, typical change of this kind.

The efficient cause also differs from the final cause, in the same sense in which the paradigmatic form of a natural thing differs from that out of which it naturally comes to be what it is. The efficient cause of a natural change potentially is this change, and the final cause is what the efficient cause potentially is: that into which it should turn when it proceeds according to the nature of the changing thing. The final cause is what a change is supposed to be; and what a change is supposed to be is not *this* particular change, but *a* change of this kind. This means that the final cause of a change may be shared by several instances of the same kind. One natural change may be what another natural change is as well. Further, the final cause of a natural change does not *occur* as such when the change occurs (cf. *Physics* B7, 198b1–4). Again, we must not confuse the essence of a particular entity with that very entity. As has been explained above, the definition of a thing should not include a reference to a particular portion of matter. In the same sense in which the particular matter of a thing does not belong to its essence, the particular time at which a change occurs does not belong to its essence.[32]

VI. CONCLUSION. AND SOME HISTORY

It should be fairly clear by now how the four Aristotelian causes can be brought into a system, and why a natural scientist must refer to all of them in her explanations, as Aristotle claims in *Physics* B7 (198a21–24). Since the proper account of a natural thing includes a principle of motion and rest, we can only understand what it is to be that kind of thing by taking into account in what ways such things typically change. Conversely, since a natural change is a change that is governed by a principle of motion and rest in the changing thing, we could not understand natural changes without also investigating the nature of the things that are involved. Further, as Plato points out in the *Timaeus* (for example, 47E), the student of nature must always ask two kinds of question about the object of her studies: "What is it for, or where does it lead to?" and "What is required or necessary?" The first question concerns what a thing or change is supposed to be and do, the second is about that which potentially satisfies such requirements. When investigating substances, the natural scientist must therefore investigate both their paradigmatic form and that which may come to realize that form, for instance their matter. When studying natural changes, the same questions take the form of the question about their final cause or typical course and the required steps that they typically involve. In each case, the two questions are complementary, just as the matter and the form of a thing complement one another. The efficient cause is for a change what the material cause is for a thing. The final cause is for a change what the formal cause is for a thing. Whereas the material cause is that which potentially is a natural thing, the paradigmatic form is what it naturally comes to be; and whereas an efficient cause is that which potentially is a natural change, the final cause is what this change naturally comes to be.

---

[32] There are, of course, events that essentially occur at certain dates, like Ramadan or Christmas. But the date that will be mentioned in their definition is a generic date. The case is analogous to the case of "functional matter" dealt with above.

The four causes may be, and have been, classified according to several different criteria; and while these criteria are sometimes similar to the ones suggested in this paper, they do not exactly match. Avicenna, for instance, distinguishes between external and internal causes. He says that the formal and material cause of a thing are included in its subsistence and are part of its existence, while the efficient and final cause are not (*Kitāb ash-Shifā'*, *Metaphysics* VI,1).[33] Avicenna's efficient cause, for instance, is an *agent*, and as such clearly distinct from what it causes. This distinction between internal and external causes may be combined with Avicenna's distinction between causes of reality and causes of existence. Since there is a distinction between a thing's existence and its reality,[34] there is also a distinction between the causes of its existence and the causes of its reality. Avicenna writes:

> In its reality, the final cause is a reason for the rest of the causes to exist as causes in action; and in its existence, the final cause is a consequence of the existence of the other causes as causes in action. It is as though the reality of the final cause is the cause of the cause of its existence, and its existence the effect of the effect of its reality. Its reality, however, is not a cause unless it comes about as an image in the soul or something similar (VI,5).[35]

This makes good sense in cases where the final cause literally is something like a goal in an agent's mind. The representation of this goal (its reality) causes the agent to bring it into existence, so that the existence of the goal state is caused by an action that is caused by the representation of this goal state. Causes of essence may thus be taken to be causes that operate via a representation or desire in an agent. For Avicenna, it seems, only agents who are capable of representing goals states can be efficient causes.[36]

---

[33] G.C. Anawati et al., eds. (Cairo: Matba'a al-Amīrīya 1960), p. 258.

[34] Reality = *shay'iyya*, from *shay'* = Latin *res*. Compare the medieval notion of *realitas objectiva* as the objective *content* of an idea.

[35] Anawati et al., eds., p. 292; my translation. The Arabic words for "actuality" (*fi'l*), "efficient" (*fā'iliya*), and "agent" (*fā'il*) derive from the same root.

[36] This is subject to several qualifications that I cannot discuss here. The word for "intention" (*ma'nan*)—which Avicenna uses in his *Kitāb al-Ishārāt wa-t-Tanbīhāt* II,4 (Forget, ed. (Leiden, 1892), p. 140)—may refer to the meaning or import of something; the word for "soul" (*nafs*) may also be used as a reflexive pronoun. Taking all this into account, one might be able to construe Avicenna as saying that all efficient causes must possess in themselves a certain directedness (instead of having an intention in their minds).

Aquinas, in his account of the four causes, follows Avicenna.[37] He claims, first, that all efficient causality necessarily involves final causality. For an efficient cause is necessarily a cause of something specific: what it causes must be of some determinate type and have a certain limit. Hence, there can be no efficient causality without final causality, just as there can be no actual matter without form. And since matter and form are relevant only when a thing changes, namely when the matter is shaped by an efficient cause such as to take on a certain form, the final cause is the most important of all causes. All this is compatible with the account of the Aristotelian doctrine given in this paper. Material and efficient causes are indeed necessarily causes *of something*, and what they are causes of are the paradigmatic form and the final cause, respectively.

However, Aquinas then goes on to argue that there can be no final causality *nisi ex intentione finis*, and that therefore, all efficient causation must be a sort of intentional action.[38] He writes:

> That which possesses reason moves toward an end by itself, because it can freely determine its own actions, by use of its will and reasoning capabilities. That however which lacks reason tends toward its end by a natural inclination. It is, as it were, moved by something else and does not move by itself. For it does not understand the reason that is provided by an end and therefore, it cannot ordain anything to an end, but is only directed toward it by something else. In this sense, the whole of nonrational nature relates to God as a tool to an agent. (*Summa Theologiae* IaIIae 1,2 c.a.)

According to Aquinas, nonrational beings cannot move by themselves, that is, they cannot efficiently cause changes. He argues, following Avicenna, that all efficient causation involves final causality, and that final causes can only act by representing goal states to intentional agents. This implies that all natural changes must ultimately go back to some intentional action, performed either by human beings or by God (angels and demons aside). In the world as Aquinas depicts it, only rational agents can be efficient causes.

This is of course only a very rough sketch of two episodes in the history of Aristotle's division of causes. It may well be too fragmentary to be accurate. Supposing it is correct, however, there are two possible reactions to Avicenna and Aquinas. First, one may argue that there are no final causes in nature. But then, by the same token, there will be no efficient causes in nature. The laws of physics will only bring events under general laws, rather than offering causal explanations in any strict sense. This is in fact what many philosophers, from Hume to Russell, have suggested. Alternatively, one may point out that since there is indeed efficient causality beyond intentional agency, there must also be *final* causality in the absence of intentional agency. That is, if there are laws according to which certain events naturally result in other events, these laws will describe their final cause.

---

[37] See *De Principiis Naturae* §3; *Summa Theologiae* IaIIae 1,2 c.a.; *Summa Contra Gentiles* III,1 §2.

[38] Again, an alternative reading is possible. As Paul Hoffman points out ("Does Efficient Causation Presuppose Final Causation? Aquinas vs. Early Modern Mechanism," (unpublished manuscript)), the *intentio finis* may simply be an inclination towards a certain end state (cf. *Summa Theologiae* IaIIae 12,1 c.a.). The passage from IaIIae 1,2 c.a. quoted above, however, clearly states that all directedness is due to a rational agent.

What one cannot do, within an Aristotelian framework, is to maintain that there are efficient causes, but no final causes in nature. This would be to sever the conceptual link between efficient and final causality, disregarding the fact that all efficient causes are causes of determinate movements and thus are necessarily complemented by final causes in the sense specified in the body of the present paper. That a cause is a cause of something specific means that it naturally turns into something specific when it occurs; and what a change naturally turns into is its final cause.

The first reaction, that there are neither final nor efficient causes in nature, has its obvious limitations when it comes to accounting for the behavior of living beings. This is of course what critics of Descartes have always emphasized, but it is rarely noticed that despite all such criticism, we remain Cartesians in that we still tend to assume that final causality must be confined to some "mental realm." The quite natural and harmless teleology that is involved when we classify natural things according to paradigmatic forms and when we expect a natural process to take a specific course has been mystified by the assumption that all teleology, not only of the kind reflected in human intentionality, is confined to the sphere of intentional actions. On the other hand, once we realize that the gap between natural things and intentional agents is not as radical as it seems, since paradigmatic forms and final causes may be found on both sides, we should have fewer problems with questions that have puzzled thinkers of the twentieth century, such as the nature of biological functions and natural kinds, or the demand for an entirely naturalistic account of the human mind. There is indeed hope for such a naturalistic account, since after all, naturalism need not defy teleology.

BORIS HENNIG

Humboldt University of Berlin
University of Pittsburgh

COMMENTS AND CRITICISM

PAINS, IMPERATIVES, AND INTENTIONALISM

The imperative theory of pain is a useful and powerful generalization of intentionalism.[1] Demonstrating that intentional content need not be declarative increases the strategies available to those who hope to account for our mental life in terms of intentional content. This comment raises two problems for Colin Klein's imperative theory. First, even if he has shown that pain essentially has imperative content, he has not fully vindicated his claim that this content exhausts pain sensations. Second, while he acknowledges that his account does not easily encompass pains like generalized headaches and visceral aches, pains of that kind pose a particular problem to the imperative account that he does not address. But his paper contains a strong argument against an account—the tissue damage declarative account of pain—that appears able to handle such pains. Intentionalism,. therefore, is more undermined than defended by his interesting paper.

I. IMPERATIVE CONTENT DOES NOT EXHAUST PAIN

Let us grant the account Klein gives of sensations with positive imperative content—such as thirst and hunger. So: a sensation of thirst demands that I drink, an itch that I scratch. A given bodily demand can be satisfied in a number of ways—drinking either water or beer satisfies the command of thirst, and scratching on bare skin or through clothing satisfies the command of itch. Klein holds that for such sensations, "[t]he distinctive qualities of each stem from the fact that they demand a certain kind of action to be satisfied" (519).

Klein makes the case that pain is like itch and thirst. He argues that all of these sensations are "often poorly localized," interfere with an agent's ability to carry out plans unrelated to the sensation's commands, and do not "carry information about their causal origins" (519). This establishes enough similarity between pain and thirst, for example, to warrant treating pains as having imperative content. Klein addresses a disanalogy: pains do not appear to command us to do anything in particular. Klein concludes that pains have negative imperative content—they issue commands not to perform certain sorts of bodily actions (520).

---

[1] Colin Klein, "An Imperative Theory of Pain," this JOURNAL, CIV, 10 (October 2007): 517–32.

0022-362X/09/0603/161–66

But Klein never addresses the most fundamental disanalogy between pain and the sensations he characterizes as having positive imperative content. In typical cases, it suffices, for a feeling of hunger or thirst to cease, that one comply with the imperative—that one eat or drink an adequate amount. But pain, even pain of the kind for which Klein's account works best, is not like this. Consider the pain in one's ankle as one "recuperate[s] from a sprain" (520). It has, Klein argues, imperative content demanding that one not put weight on the ankle, that one not rotate the ankle, and so on—that one not move or use one's ankle in ways that would further injure it. Klein is right that the phenomenology of pain varies with attention and plans for action. He is also right that many pains lessen or disappear entirely when one obeys their relevant imperatives (after all, some muscle and ligament pain appears only upon motion). But many pains do not, even pains in muscles or joints. Many pains persist, and persist at the same level of intensity, even when one obeys their commands. Thirst is simply not like this.

Not grappling with this disanology weakens the case Klein makes for the exhaustiveness of imperative content—the claim that "there is no difference in the phenomenology of a pain experience without some difference in the intentional content of a pain experience" (522). Klein writes:

> The difference between someone in pain and someone not in pain is just the presence or absence of an imperative. The pain of a broken ankle is an imperative for me not to use my ankle. Give me enough morphine for my pain to be lessened, and you lessen the strength of that imperative. Give me enough to eliminate my pain, and you have eliminated the imperative against action entirely (522).

The variation in imperative strength that is supposed to suffice for variation in pain experience is not caused here by an agent's obeying the imperative. Imperatives are usually withdrawn when they are obeyed. But they can also be removed without being obeyed. And it is telling that to produce a case where an imperative—and so pain, on his account—disappears entirely, Klein has to talk about removal rather than obedience. I can entirely remove an imperative to drink if I take a drug that interferes with sensations of thirst. I can also remove an imperative to drink by obeying it—by drinking to satiety. But perfect compliance with pain's imperatives rarely results in the withdrawal of the imperative.[2] Only the suppression or destruction of the imperative, by an effective analgesic, does that.

---

[2] Separately from the line of argument that follows, one might worry that content that is not withdrawn when (what had appeared to be) its fulfillment conditions are

The fact that the imperative—and so the pain—outlasts our obedience is unfortunate for us.[3] It does not by itself show that imperative content fails to exhaust pain. But when viewed in connection with another commitment of Klein's, it supports the case for nonexhaustion. Klein argues that pain can only serve its biologically protective role if the connection between a pain and its proscribed actions is noncontingent. He argues that any merely declarative content—for example, what standard tissue-damage intentionalism attributes to pain—will be only contingently linked with the actions that are required for the body's protection. Klein asserts that no merely declarative content, no representation of a state of affairs in the world, can ever necessitate any particular response.[4] Only an imperative can do that.

In arguing against tissue-damage representationalism, Klein commits himself to noncontingent links between pains and prohibited actions. This seems not only biologically but phenomenologically sound. Pain, and the positive imperative sensations generally, seem from the inside to be linked to what is good for us when we are in the conditions linked to those sensations. (Drinking is what a creature whose condition gives rise to a sensation of thirst needs to do, and that is what it feels like to the thirsty creature.) But the aim of accounting entirely for differences between sensations (for example, between thirst and pain) and between kinds of pains in terms of the different imperatives they issue is complicated by the fact that pain-cessation is not a matter of imperative-obedience. It is not plausible that pain should be so tightly connected to prohibited action as to be constituted necessarily by the imperatives to avoid just those actions, and yet persist with exactly the same phenomenological content in the face of perfect obedience to those very imperatives. Klein needs a reason why negative but not positive imperatives persist in the face of

---

met cannot be imperative content. In contrast, even thought it may violate pragmatic rules to repeat declarative announcements after they have been understood, such violations do not undermine the content's status as declarative.

[3] Klein might argue that the unignorable aversiveness of pain is necessary if it is to serve any protective function, and assert that imperative-persistence is an unfortunate consequence of any adequately strong warning mechanism. But then he needs to explain why warnings he counts as equally important, such as thirst (see p. 526), do not persist after obedience. He could try to appeal to his suggestion that pains that persist after damage have healed (because they obviously cannot be reporting on present damage) concern "weakness and propensity to re-injury" (523). But while that could explain an imperative that re-emerged with fresh motion, it does not answer the question of why a given episode of pain persists after and while its particular imperative is obeyed.

[4] See p. 526 for these arguments that the "imperative sensations generally have the feature that informative content (and the resultant possibility of contingency of response) would be maladaptive."

perfect obedience. He needs to explain why it is only in the case of positive imperative sensations that satisfying their distinctive imperatives satisfies the person whose sensations they are.

Claiming that this just is the difference between positive and negative imperatives might be a good strategy. But even using this strategy, Klein would have to develop his account of how pains are distinguished from one another, because he will not be able to appeal (as he could in the case of the positive sensations) to both the contents of their imperatives and their distinctive conditions of full satisfaction. Pains all have, it seems, the same condition of full satisfaction—the administration of an analgesic or (local or general) anesthetic. The imperative account may be able to explain why negative imperative sensations are not as obedience-responsive as positive imperative sensations are, and do the work of distinguishing among distinct negative imperative sensations with only one of the two tools available to accounts of positive imperative sensations. But until it does, it remains plausible to suppose that part of pain experience is not exhausted by imperative content. That is the simplest explanation of why pain persists after and during obedience to its imperatives.[5]

II. PAINS IN THE GUT

Klein gives a list of pains that, he claims, pose problems for all accounts: those pains "associated with headache, menstrual cramps, and disturbances of the deep viscera" (530). He notes that they pose a particular problem for his account because "they do not appear to have the tight connection to action that most pains have" (530). While leaving the details to future work, he hopes that the imperative theory could treat these pains as "proscrib[ing] against large movements of the head or torso," and avoid having to treat them as "degenerate cases: cases where pain is proscribing against unspecified or unintelligible actions" (530).[6]

---

[5] Klein's discussion of morphine pain (527–30) does not provide direct help here. Establishing that negative imperatives (against new action) could still be issued to someone who claims not to care about his pain does not eliminate the question of why imperatives persist after the perfect obedience of someone who very much cares about his pain.

[6] There are presumably several reasons why Klein wishes to avoid this. One factor is the importance of ensuring that pain's intentional content carries imperative force in a serious sense. Without this, Klein's account could slide into the tissue-damage declarative intentionalism he criticizes. Arguably, a declarative content that something is damaged implies that it ought to be fixed. But 'Fix this damage!' or 'Remove the cause of this damage!', while they arguably could be contents for almost every pain, are hardly the kind of contents that could explain phenomenal differences between pain, or that could serve an adaptive purpose. For both these reasons, Klein needs the actions proscribed by pain's intentional content to be specific and intelligible. I am grateful to an anonymous referee for this JOURNAL for pressing me to consider this issue.

Klein may be right that the pain of headache involves negative imperatives not to move the head or neck—certainly headaches seem often to require total rest, and resting is a good way to avoid moving one's head. But he faces something worse than his worst case scenario for some pains in the torso. The pains associated with menstruation, early and middle obstetric labor (where there is no felt urge to push), and such varied events as bladder spasms and intestinal cramps are unique because they result from bodily activity that will continue regardless of whether the embodied agent wants it to or not. Here, the problem is not (only) that pain may persist after a negative imperative is obeyed.[7] Rather, the problem is that it is not clear what the relevant imperative would be. There is no way to stop uterine contractions. (They can be slowed, sometimes to the point of cessation, pharmacologically. But 'Take this drug!' is both a positive imperative and a poor fit with the evolutionary account Klein wants to give of the protective function of pain.) Perhaps 'Keep contracting!' is the unintelligible action proscribed by uterine pain, on the worst-case scenario Klein envisions. But the problem seems to go deeper.

There is nothing unspecified or unintelligible about the action of uterine contraction. The problem is that an imperative to avoid (or perform) that action cannot sensibly be understood as addressed to the agent. Certainly, it is not intuitively appropriate in anything like the way that an imperative proscribing ankle rotation is. Significantly many of the pains—including perhaps cardiac pain—whose felt locations are somewhere in the torso appear to have this feature. (The pains of breathing with a cracked rib or severe asthma are exceptions. But 'Do not breathe!' is a problematic imperative for other reasons.)

Klein is right that there are many pains whose phenomenology is deeply connected to an agent's possibilities for action. But this is not so for pains in the torso. One of the more salient facts about the phenomenology of these pains is how tightly connected they are to activities of what feel like (parts of) the body in contradistinction to the self. (This is true even if the embodied self is able to alter significantly the secondary affect of, for example, labor pain by viewing it not as a sign of distress, damage, or disease, but as part of a project she endorses.) Whatever its other problems, a declarative account does not have a special difficulty with such pains. If pains have intentional content representing tissue damage, they can have such content no matter where or why such damage is occurring.[8] Given how common pains of this type are, the imperative theory needs a way to handle them. Otherwise it is only an account of certain kinds of pain, and not of the features common to every pain experience.

---

[7] Though that is a problem. It will surely be difficult to come up with different torso movements for ulcer pain and menstrual pain to proscribe, and hence difficult to make good the claim that the phenomenological difference between the two is exhausted by the difference between their respective imperatives. And if that can be done, it becomes once again unclear why perfect obedience does not produce complete pain relief.

III. CONCLUSION

The imperative theory of pain faces two difficulties the declarative theory does not: it needs to explain why pain persists after and while its imperatives are obeyed, and it needs to explain how to handle pains connected to bodily actions not subject to intentional control. But Klein's criticisms of the standard declarative view—tissue damage representationalism—are deeply damaging. Unless the imperative account is improved, or there are prospects for pain-contents that are neither exclusively imperative nor exclusively declarative, intentionalism about pain appears to be in trouble.

MAURA TUMULTY

Colgate University

---

[8] Pain due to uterine contractions would, however, be an additional example Klein could use in support of his argument that "[m]any pains show up in nondamage situations" (523). Michael Tye, however, is perfectly willing to accept that pain can occur without tissue damage ("In Defense of Representationalism: Reply to Commentaries," in Murat Aydede, ed., *Pain: New Essays on Its Nature and the Methodology of Its Study*, (Cambridge: MIT, 2005, pp. 163–76, see p. 166)). He even holds that pain—understood as a quality experienced in having a pain experience—is not "tissue damage simpliciter," but tissue damage represented by a token pain experience of a person at a time ("Reply," p. 166). He accepts that pain (what is experienced in pain experience) "has the status of an 'intentional inexistent'" (*ibid.*). Giving this nonobjective status to pain-as-an-object-of-experience is in tension with Tye's claim that pain represents simply by tracking tissue damage ("Another Look at Representationalism about Pain," in *Pain: New Essays on Its Nature and the Methodology of Its Study*, pp. 99–120, on p. 112). But it does expand Tye's ability to insist that he has no special problem with pain experience in the absence of actual tissue damage.

THE JOURNAL OF PHILOSOPHY

Special Issues Available

Parts and Wholes — December 2006
On Bernard Williams — June 2003
Causation — April 2000
Jürgen Habermas and John Rawls — March 1995

The Dewey Lectures:
Michael Dummett — *Truth and the Past*, 2002
Hilary Putnam — *Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind*, 1994
Donald Davidson — *The Structure and Nature of Truth*, 1990
Amartya Sen — *Well-being, Agency, and Freedom*, 1985
John Rawls — *Kantian Constructivism in Moral Theory*, 1980
Ernest Nagel — *Teleology Revisited*, 1977
W.V. Quine — *Ontological Relativity*, 1968

The Woodbridge Lectures:
John McDowell – *Having the World in View: Sellars, Kant, and Intentionality*, 1997

Subscriptions

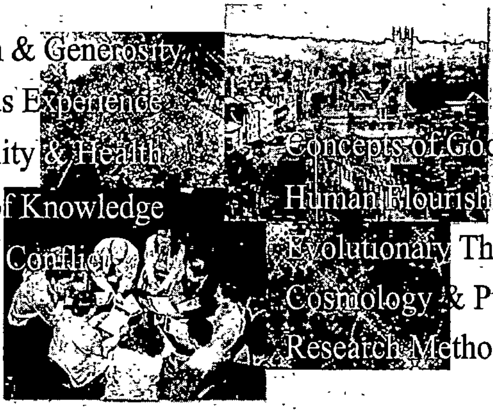
Individuals $45.00
Students $20.00
Libraries and Institutions $100.00
Print and Online for Libraries and Institutions $180.00
Postage outside the United States $15.00

Special Offer

One complimentary issue of any of the above issues with each new subscription.
All issues are $5.00 plus $3.00 for shipping and handling.
Send check or money order to:
The Journal of Philosophy
Columbia University
1150 Amsterdam Avenue, Mail Code 4972
New York, NY 10027

---

www.journalofphilosophy.org



$30,000
MGNI GRANTS AVAILABLE

For *transdisciplinary* exploration of areas such as:

Altruism & Generosity
Religious Experience
Spirituality & Health
Nature of Knowledge
Peace & Conflict
Concepts of God
Human Flourishing
Evolutionary Theory
Cosmology & Purpose
Research Methodology

and much more...

Proposal deadline for upcoming grant round is
30 September 2009.

Metanexus
GLOBAL
NETWORK
INITIATIVE

To learn more about Metanexus
Institute and its programs
join us at our 2009 conference
Cosmos, Nature, and Culture
18-21 July in Phoenix, AZ
Or visit us online at
www.metanexus.net/jop

www.metanexus.net/jop

he review of

metaphysics

philosophical quarterly

SN 0034-6632

arch 2009 | **VOL. LXII, No. 3** | **ISSUE No. 246** | **$15.00**

ticles

ADAM SEAGRAVE	*Cicero, Aquinas, and Contemporary Issues in Natural Law Theory*
ANIEL GUEVARA	*The Will as Practical Reason and the Problem of Akrasia*
HRISTOPHER JOHNS	*The Grounds of Right and Obligation in Leibniz and Hobbes*
ICHOLAS RESCHER	*Fallacies Regarding Free Will*
ENN GOODMAN	*An Idea Is Not Something Mute Like a Picture on a Pad*
RANDON ZIMMERMAN AND STAFF	*Summaries and Comments*

book reviews ***philosophical abstracts*** ***announcements***

ndividual Subscriptions $40 Institution Subscriptions $60 Retired/Student Subscription $25

VISA, MASTERCARD, AMEX AND DISCOVER ACCEPTED
DISCOUNTS AVAILABLE ON TWO AND THREE YEAR SUBSCRIPTIONS

FOR MORE INFORMATION, AND TO SUBSCRIBE OR RENEW ONLINE, GO TO:

www.reviewofmetaphysics.org

Jude P. Dougherty, Editor
The Catholic University of America, Washington D.C. 20064
email: mail@reviewofmetaphysics.org

FOUNDATIONAL ADVENTURES

Conference in Honor of the 60th Birthday of Harvey M. Friedman
The Blackwell Inn, The Ohio State University, Columbus, OH 43210
May 14-17, 2009

Conference webpage:
http://people.cohums.ohio-state.edu/tennant9/friedman_conference.html

This interdisciplinary conference, with a stellar list of speakers, is made possible by the support of The Ohio State University and the John Templeton Foundation.

We have at our disposal some limited travel funds from the National Science Foundation, with the aim of providing opportunities for participation by women, minorities, and persons with disabilities. Details about the application process are available on the conference webpage.

The program and the list of titles and abstracts are also accessible from that webpage.

Plenary speakers include Martin Davis (UC Berkeley and NYU), Solomon Feferman (Stanford), Harvey Friedman (Ohio State), Saul Kripke (CUNY), Anil Nerode (Cornell), Hilary Putnam (Harvard), Gerald Sacks (Harvard and MIT) and Patrick Suppes (Stanford).

You can now subscribe to the *Journal of Philosophy* using

PayPal

To learn more about this easy new option, click on the PayPal icon at **www.jphil.org.**

REVUE
INTERNATIONALE DE
PHILOSOPHIE

Direction: Michel Meyer

c / o Université Libre de Bruxelles, C.P. 188
50, avenue Franklin-Roosevelt / B-1050 Bruxelles (Belgique)
Tél. 32 2 650 25 61 / Fax 32 2 650 45 61
e-mail : tsalis@ulb.ac.be

1-2009 / volume 63 / n° 247

Shakespeare

Agnes HELLER
Preface

Paul KOTTMAN
Disinheriting the Globe: On Hamlet's Fate

Peter MURPHY
The Power and the Imagination:
The Enigmatic State in Shakespeare's English History Plays

David ROBERTS
Likeness and Difference: The Logic of Plot in *Measure for Measure*

Zoltán GYENGE
The Heteronomy of sin. Richard, Agamemnon or perhaps Abraham?

Prochains fascicules / Forthcoming

Philosophie du roman
Goethe
Hintikka
Le théâtre des Grecs au romantisme

The MIT Press

THE MONSTROSITY OF CHRIST

Paradox or Dialectic?

Slavoj Žižek and John Milbank
edited by Creston Davis

A militant Marxist atheist and a "Radical Orthodox" Christian theologian square off on everything from the meaning of theology and Christ to the war machine of corporate mafia.

Short Circuits series, edited by Slavoj Žižek • 416 pp., $27.95 cloth

CONSCIOUSNESS REVISITED

Materialism without Phenomenal Concepts

Michael Tye

"This marvelously informed, powerfully argued book is Michael Tye's latest contribution to the task of finding a naturalistic understanding of consciousness."
— Frank Jackson, Princeton University

Representation and Mind series • 256 pp., 17 illus., $35 cloth

THE CRUCIBLE OF CONSCIOUSNESS

An Integrated Theory of Mind and Brain

Zoltan Torey
foreword by Daniel C. Dennett

"Just as poets often find that the constraints of rhyme and meter force them to discover strikingly apt expressions of their thoughts, it turns out that couching a computational theory of the mind in resolutely noncomputational terms pays dividends."
— Daniel C. Dennett, Tufts University and author of *Sweet Dreams*

272 pp., 2 illus., $34 paper

DISJUNCTIVISM

Contemporary Readings

edited by Alex Byrne and Heather Logue

"Disjunctivism is probably the single most important idea in philosophy of perception, epistemology, and theory of reference today. I don't know a better place to begin than with the cool, lucid overview provided by the editors, itself a masterly essay. An indispensable collection." — John Campbell, University of California, Berkeley

368 pp., $36 paper

BEYOND RED AND BLUE

How Twelve Political Philosophies Shape American Debates

Peter S. Wenz

"*Beyond Red and Blue* tackles tough issues from euthanasia to torture to global trade. Peter Wenz's conclusions may make you cheer or curse, but they are sure to make you think."
— Senator Dick Durbin, Illinois

376 pp., $27.95 cloth

CONCEPTUAL ANALYSIS AND PHILOSOPHICAL NATURALISM

edited by David Braddon-Mitchell
and Robert Nola

"In this excellent volume, [the editors] have brought together new essays by the most important contributors to the debate about the viability of the Canberra Plan, both defenders and skeptics. Anyone with a serious interest in metaphysics, a priori knowledge, the philosophy of mind, and meta-ethics will need to possess this book." — Tim Crane, University College London

A Bradford Book • 384 pp., $38 paper

THE TIME OF OUR LIVES

A Critical History of Temporality

David Couzens Hoy

"Drawing on years of immersion in the major traditions of continental philosophy, Hoy has produced a remarkable meditation on the implications of each for the perennially vexed question of temporality." — Martin Jay, University of California, Berkeley

328 pp., $34 cloth

PREDICATIVE MINDS

The Social Ontogeny of Propositional Thinking

Radu J. Bogdan

"Bogdan presents a highly original synthesis that draws on recent work in the philosophy of language, developmental psychology, evolutionary biology, and the philosophy of mind. This deserves to be an influential book."
— Robert M Gordon, University of Missouri, St. Louis

184 pp., 3 illus., $25 cloth

To order call 800-405-1619 • **http://mitpress.mit.edu**

The MIT Press

FUNCTIONS IN BIOLOGICAL AND ARTIFICIAL WORLDS

Comparative Philosophical Perspectives

edited by Ulrich Krohs and Peter Kroes

"The importance of [the book] extends beyond philosophy of biology and philosophy of technology into issues of general philosophical interest. Highly recommended." — Lynne Rudder Baker, University of Massachusetts Amherst

Vienna Series in Theoretical Biology • 312 pp., 12 illus., $50 cloth

THE ETHICS OF PROTOCELLS

Moral and Social Implications of Creating Life in the Laboratory

edited by Mark A. Bedau and Emily C. Parke

Experts explore the potential benefits, risks, and moral aspects of protocell technology, which creates simple forms of life from nonliving material.

Basic Bioethics series • 392 pp., $28 paper

COGENT SCIENCE IN CONTEXT

The Science Wars, Argumentation Theory, and Habermas

William Rehg

"Rehg's riveting account of cogency will revolutionize the way critical theorists, logicians, and rhetoricians think about argumentation in general and, more particularly, about how we should assess the social institutionalization of collaborative research and debate in science proper for years to come." — David Ingram, Loyola University, Chicago

Studies in Contemporary German Social Thought
360 pp., 4 illus., $40 cloth

ZONE BOOKS

Now in Paper

THE CRADLE OF HUMANITY

Prehistoric Art and Culture

Georges Bataille

edited and introduced by Stuart Kendall
translated by Michelle and Stuart Kendall

A radically interdisciplinary inquiry into the origins of human consciousness, community, and potential.

Distributed for Zone Books • 224 pp., 15 illus., $19.95 paper.

PHILOSOPHY OF LOVE

A Partial Summing-Up

Irving Singer

The author of the classic philosophical treatment of love reflects on the trajectory, over decades, of his thoughts on love and other topics.

The Irving Singer Library • 144 pp., $14.95 cloth

Now in Paper

THE PARALLAX VIEW

Slavoj Žižek

"No one demonstrates the continued philosophical vitality of Marxism better than Slavoj Žižek." — *Tikkun*

Short Circuits series, edited by Slavoj Žižek • 448 pp., $14.95 paper

Now in Paper

THE REALLY HARD PROBLEM

Meaning in a Material World

Owen Flanagan

"The book sparkles with thought and a likeable humour." — Steven Poole, *The Guardian*

304 pp., 1 illus., $15.95 paper

Now in Paper

NEUROPHILOSOPHY OF FREE WILL

From Libertarian Illusions to a Concept of Natural Autonomy

Henrik Walter

translated by Cynthia Klohr

"Walter uses contemporary findings in neurobiology and new understandings in the foundations of nature to explore 'free will.' The result is a fresh thesis well worth considering." — Kevin T. French, *Academia*

408 pp., 7 illus., $27 paper

Now in Paper

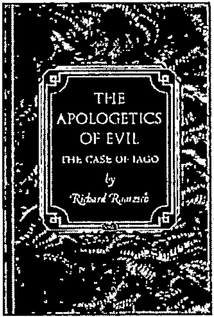
RATIONALITY AND LOGIC

Robert Hanna

"This stimulating and wide-ranging book will be of interest to philosophers of logic and also to cognitive psychologists." — Jane Heal, University of Cambridge

344 pp., $18 paper

To order call 800-405-1619 • **http://mitpress.mit.edu**



The Apologetics of Evil

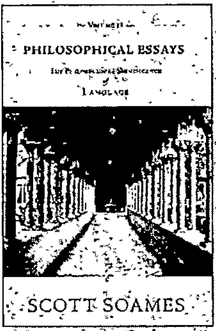
The Case of Iago

Richard Raatzsch

Translated by Ladislaus Löb

"What is evil? What are its forms? How is it motivated? Raatzsch's treatment of [these questions] is sensitive, imaginative, and broadly based. This book brings together different lines of argument from epistemology, moral philosophy, and philosophy of mind in a highly compact and supercharged, yet fully comprehensible form. The result packs an enormous intellectual punch."
—Raymond Geuss, University of Cambridge

Princeton Monographs in Philosophy
Harry G. Frankfurt, Series Editor
Cloth $26.95 978-0-691-13733-9



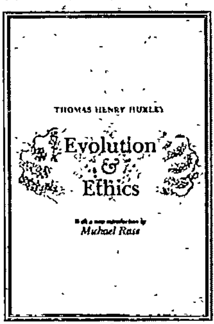
Philosophical Essays, Volume 2

The Philosophical Significance of Language

Scott Soames

The two volumes of *Philosophical Essays* bring together the most important essays written by one of the world's foremost philosophers of language. The essays in Volume 2 illustrate the significance of linguistic concerns for a broad range of philosophical topics—including the relationship between language and thought; the distinction between metaphysical and epistemic possibility; the nature of necessity, actuality, and possible worlds; and skepticism about meaning and mind.

Paper $39.50 978-0-691-13683-7
Cloth $95.00 978-0-691-13682-0



Evolution and Ethics

Thomas Henry Huxley

Evolution and Ethics, widely considered to be Thomas Henry Huxley's greatest lecture, distilled a lifetime's wisdom and sensitive understanding of the nature and needs of humankind. This new edition features an introduction by renowned historian and philosopher of biology Michael Ruse, placing Huxley's lecture in its original context while showing its even deeper relevance for our own time.

Paper $16.95 978-0-691-14130-5
Cloth $39.50 978-0-691-14192-3

PRINCETON UNIVERSITY PRESS press.princeton.edu

CENTRAL LIBRARY
UNIVERSITY
OF
CALCUTTA

THE JOURNAL OF PHILOSOPHY

VOLUME CVI, NUMBER 4
APRIL 2009

SPECIAL ISSUE
OUR KNOWLEDGE OF NATURE AND NUMBER: GROUNDS AND LIMITS

Edited by Carol Rovane

page
167 *The Mysteries of Nature: How Deeply Hidden?* Noam Chomsky
201 *Gödel, Nagel, Minds, and Machines* Solomon Feferman
220 *Where Our Number Concepts Come From* Susan Carey

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane; Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Michael J. Mooney; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2009

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.
POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2009 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 4, APRIL 2009

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA
ALIPORE CAMPUS

THE MYSTERIES OF NATURE: HOW DEEPLY HIDDEN?

The title for these remarks is drawn from Hume's observations about the man he called "the greatest and rarest genius that ever arose for the ornament and instruction of the species," Isaac Newton. In Hume's judgment, Newton's greatest achievement was that while he "seemed to draw the veil from some of the mysteries of nature, he shewed at the same time the imperfections of the mechanical philosophy; and thereby restored [Nature's] ultimate secrets to that obscurity, in which they ever did and ever will remain." On different grounds, others reached similar conclusions. Locke, for example, had observed that motion has effects "which we can in no way conceive motion able to produce"—as Newton had in fact demonstrated shortly before. Since we remain in "incurable ignorance of what we desire to know" about matter and its effects, Locke concluded, no "science of bodies [is] within our reach" and we can only appeal to "the arbitrary determination of that All-wise Agent who has made them to be, and to operate as they do, in a way wholly above our weak understandings to conceive."[1]

I think it is worth attending to such conclusions, the reasons for them, their aftermath, and what that history suggests about current concerns and inquiries in philosophy of mind.

The mechanical philosophy that Newton undermined is based on our common-sense understanding of the nature and interactions of objects, in large part genetically determined, and, it appears, reflexively yielding such perceived properties as persistence of objects through time and space, and as a corollary their cohesion and conti-

---

[1] David Hume, *The History of England*, VI, LXXI. John Locke, *An Essay concerning Human Understanding*, Book IV, chapter III. Locke's reasons, of course, were not Hume's, but relied on the boundaries of "the simple ideas we receive from sensation and reflection," which prevent us from comprehending the nature of body or mind (spirit).

0022-362X/09/0604/167–200

nuity[2]; and causality through contact, a fundamental feature of intuitive physics, "[b]ody, as far as we can conceive, being able only to strike and affect body, and motion, according to the utmost reach of our ideas, being able to produce nothing but motion," as Locke plausibly characterized common-sense understanding of the world—the limits of our "ideas," in his sense. The theoretical counterpart was the materialist conception of the world that animated the seventeenth century scientific revolution, the conception of the world as a machine, simply a far grander version of the automata that stimulated the imagination of thinkers of the time much in the way programmed computers do today: the remarkable clocks, the artifacts constructed by master artisans like Jacques de Vaucanson that imitated animal behavior and internal functions like digestion, the hydraulically activated machines that played instruments and pronounced words when triggered by visitors walking through the royal gardens. The mechanical philosophy aimed to dispense with forms flitting through the air, sympathies and antipathies, and other occult ideas, and to keep to what is firmly grounded in common-sense understanding, and intelligible to it. As is well known, Descartes claimed to have explained the phenomena of the material world in mechanistic terms, while also demonstrating that the mechanical philosophy is not all-encompassing, not reaching to the domain of mind—again pretty much in accord with the common-sense dualistic interpretation of oneself and the world around us.

I.B. Cohen observes that "there is testimony aplenty in Newton's *Principia* and *Opticks* to his general adherence to the Cartesian mechanical philosophy."[3] The word "general" is important. Newton was much influenced by the neo-Platonic and alchemical traditions, and also by the disturbing consequences of his own inquiries. For such reasons he sometimes modified the more strict Cartesian dichotomy of matter and spirit, including in the latter category "the natural agencies responsible for the 'violent' motions of chemical and electrical action and even, perhaps, for accelerated motion in general," Ernan McMullin shows in a careful analysis of the evolution of Newton's struggle with the paradoxes and conundrums he sought to resolve. In Newton's own words, "spirit" may be the cause of all movement in nature, including the "power of moving our body by our thoughts" and "the same power in other living creatures, [though] how this is done and by what laws we do not know. We cannot say that all nature is not alive."[4]

---

[2] R. Baillargeon, "Innate Ideas Revisited: For a Principle of Persistence in Infants' Physical Reasoning," *Perspectives on Psychological Science*, III (2008): 2–13.

[3] Cohen, *Revolution in Science* (Cambridge: Harvard, 1985), p. 155.

Going a step beyond, Locke added that we cannot say that nature does not think. In the formulation that has come down through history as "Locke's suggestion," he writes that "Whether Matter may not be made by God to think is more than man can know. For I see no contradiction in it, that the first Eternal thinking Being, or Omnipotent Spirit, should, if he pleased, give to certain systems of created senseless matter, put together as he thinks fit, some degrees of sense, perception, and thought." Furthermore, just as God had added inconceivable effects to motion, it is "not much more remote from our comprehension to conceive that GOD can, if he pleases, superadd to matter a faculty of thinking, than that he should superadd to it another substance with a faculty of thinking." There is no warrant, then, for postulating a second substance whose essence is thought. And elsewhere, it "involves no contradiction [that God should] give to some parcels of matter, disposed as he thinks fit, a power of thinking and moving [which] might properly be called spirits, in contradistinction to unthinking matter," a view that he finds "repugnant to the *idea* of senseless matter" but that we cannot reject, given our incurable ignorance and the limits of our ideas (cognitive capacities). Having no intelligible concept of "matter" (body, and so on), we cannot dismiss the possibility of living or thinking matter, particularly after Newton undermined common-sense understanding.[5]

Locke's suggestion was taken up through the eighteenth century, culminating in the important work of Joseph Priestley, to which we return. Hume, in the *Treatise,* reached the conclusion that "motion may be, and actually is, the cause of thought and perception," rejecting familiar arguments about absolute difference in kind and divisibility on the general grounds that "we are never sensible of any connexion betwixt causes and effects, and that 'tis only by our experience of their constant conjunction, we can arrive at any knowledge of this relation." In one or another form, it came to be recognized that since "thought, which is produced in the brain, cannot exist if this organ is wanting," and there is no longer a reason to question the thesis of thinking matter, "It is necessary to consider the brain as a special organ designed especially to produce [thought], as the stomach and the intestines are designed to operate the digestion, the liver to filter bile," and so on through the bodily organs. Just as foods enter the stomach and leave it with "new qualities, [so] impressions arrive at the brain, through the nerves; they are then isolated and without coherence. The organ enters into action; it acts on them, and soon it sends them back changed into ideas, which the language of physiognomy and gesture, or the signs of speech and writing, manifest outwardly. We conclude then, with the same certainty, that the brain digests, as it were, the impressions, i.e., that organically it makes the secretion of thought."[6] As Darwin put the matter succinctly, "Why is thought, being a secretion of the brain, more wonderful than gravity, a property of matter?"[7]

---

[4] McMullin, *Newton on Matter and Activity* (Notre Dame: University Press, 1978), pp. 52ff. He concludes that because of Newton's vacillation in use of the terms "mechanical," "spirit," and others, it is "misleading ... to take Newton to be an exponent of the 'mechanical philosophy'" (p. 73).

[5] See Locke, *op. cit.*; and correspondence with Stillingfleet, cited by Ben Lazare Mijuskovic, *The Achilles of Rationalist Arguments* (The Hague: Martinus Nijhoff, 1974), p. 73. On the development of "Locke's suggestion" through the eighteenth century, culminating in Joseph Priestley's important work (discussed below), see John Yolton, *Thinking Matter* (Minneapolis: Minnesota UP, 1983).

Qualifications aside, Newton did generally adhere to the mechanical philosophy, but also showed its "imperfections," in fact demolished it, though to the end of his life he sought to find some way to account for the mystical principle of action at a distance that he was compelled to invoke to account for the most elementary phenomena of nature. Perhaps, he thought, there might be "a most subtle spirit which pervades and lies hid in all gross bodies," which will somehow yield a physical account of attraction and cohesion and offer some hope of rescuing an intelligible picture of the world.[8]

We should not lightly ignore the concerns of "the greatest and rarest genius that ever arose for the ornament and instruction of the species," or of Galileo and Descartes, or Locke and Hume. Or of Newton's most respected scientific contemporaries, who "unequivocally blamed [Newton] for leading science back into erroneous ways which it seemed to have definitely abandoned," E.J. Dijksterhuis writes in the classic study of the mechanistic world picture and its collapse as a substantive doctrine. Christiaan Huygens described Newton's principle of attraction as an "absurdity." Leibniz argued that Newton was reintroducing occult ideas similar to the sympathies and antipathies of the much-ridiculed scholastic science, and was offering no *physical* explanations for phenomena of the material world.[9]

---

[6] Pierre-Jean-George Cabanis, *On the Relations between the Physical and Moral Aspects of Man*, Volume I (1802) (Baltimore: Johns Hopkins, 1981).

[7] Darwin cited by V.S. Ramachandran and Sandra Blakeslee, *Phantoms in the Brain* (New York: William Morrow, 1998), p. 227.

[8] Newton, *Principia*, General Scholium.

[9] Dijksterhuis, *The Mechanization of the World Picture* (New York: Oxford, 1961; reprint, Princeton: University Press, 1986), pp. 479–80.

Newton largely agreed with his scientific contemporaries. He wrote that the notion of action at a distance is "inconceivable." It is "so great an Absurdity, that I believe no Man who has in philosophical matters a competent Faculty of thinking, can ever fall into it."[10] By invoking it, we concede that we do not understand the phenomena of the material world. As McMullin observes, "By 'understand' Newton still meant what his critics meant: 'understand in mechanical terms of contact action'."[11]

To take a contemporary analogue, the absurd notion of action at a distance is as inconceivable as the idea that "mental states are states of the brain," a proposal "we do not really understand [because] we are still unable to form a conception of *how* consciousness arises in matter, even if we are certain that it does."[12] Similarly, Newton was unable to form a conception of how the simplest phenomena of nature could arise in matter—and they did not, given his conception of matter, the natural theoretical version of common-sense understanding. Locke and others agreed, and Hume carried that failure of conceivability a long step beyond by concluding that Newton had restored these ultimate secrets of nature "to that obscurity, in which they ever did and ever will remain"—a stand that we may interpret, naturalistically, as a speculation about the limits of human cognitive capacities. In the light of history, there seems to be little reason to be concerned about the inconceivability of relating mind to brain, or about conceivability altogether, at least in inquiry into the nature of the world. Nor is there any reason for qualms about an "explanatory gap" between *the physical* and consciousness, beyond the unification concerns that arise throughout efforts to understand the world. And unless *the physical* is given some new post-Newtonian sense, there is even less reason for qualms about an "explanatory gap" than in cases where there is some clear sense to the assumed reduction base. The most extreme of such concerns, and perhaps the most significant for the subsequent development of the sciences, is the explanatory gap that Newton unearthed and left unresolved, possibly a permanent mystery for humans, as Hume conjectured.[13]

Science of course did not end with the collapse of the notion of body (material, physical, and so on). Rather, it was reconstituted in a radically new way, with questions of conceivability and intelligibility dismissed as demonstrating nothing except about human cognitive capacities, though that conclusion has taken a long time to become firmly established. Later stages of science introduced more "absurdities." The legitimacy of the steps is determined by criteria of depth of explanation and empirical support, not conceivability and intelligibility of the world that is depicted.

---

[10] *Ibid.*, p. 488; Letter to Bentley, 1693.

[11] See McMullin, *op. cit.*, chapter 3, for more detailed analysis.

[12] Thomas Nagel, *Other Minds* (New York: Oxford, 1995), p. 106.

[13] For varying perspectives on the "explanatory gap," see essays in Galen Strawson et al., *Consciousness and Its Place in Nature*, Anthony Freeman, ed. (Charlottesville, VA: Imprint Academic, 2006).

Thomas Kuhn suggests that "It does not, I think, misrepresent Newton's intentions as a scientist to maintain that he wished to write a *Principles of Philosophy* like Descartes [that is, true science] but that his inability to explain gravity forced him to restrict his subject to the *Mathematical Principles of Natural Philosophy*, [which] did not even pretend to explain why the universe runs as it does," leaving the question in obscurity. For such reasons "It was 40 years before Newtonian physics firmly supplanted Cartesian physics, even in British universities," and some of the ablest physicists of the eighteenth century continued to seek a mechanical-corpuscular explanation of gravity—that is, what they took to be a *physical* explanation—as Newton did himself. In later years positivists reproached all sides of the debates "for their foolishness in clothing the mathematical formalism [of physical theory] with the 'gay garment' of a physical interpretation," a concept that had lost substantive meaning.[14]

Newton's famous phrase "I frame no hypotheses" appears in this context: recognizing that he had been unable to discover the *physical* cause of gravity, he left the question open. He adds that "to us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea." But while agreeing that his proposals were so absurd that no serious scientist could accept them, he defended himself from the charge that he was reverting to the mysticism of the Aristotelians. His principles, he argued, were not occult: "their causes only are occult"; or, he hoped, were yet to be discovered in physical terms, meaning mechanical terms. To derive general principles inductively from phenomena, he continued, "and afterwards to tell us how the properties of actions of all corporeal things follow from those manifest principles, would be a very great step in philosophy, though the causes of these principles were not yet discovered."[15]

---

[14] Kuhn, *The Copernican Revolution* (New York: Random House, 1957), p. 259. Heinrich Hertz, cited by McMullin, *op. cit.*, p. 124.

[15] Dijksterhuis, *op. cit.*, p. 489.

To paraphrase with regard to the contemporary analogue I mentioned, it "would be a very great step in science to account for mental aspects of the world in terms of manifest principles even if the causes of these principles were not yet discovered"—or to put the matter more appropriately, even if unification with other aspects of science had not been achieved. To learn more about mental aspects of the world—or chemical or electrical or other aspects—we should try to discover "manifest principles" that partially explain them, though their causes remain disconnected from what we take to be more fundamental aspects of science. The gap might have many reasons, among them, as has repeatedly been discovered, that the presumed reduction base was misconceived, including core physics.

Historians of science have recognized that Newton's reluctant intellectual moves set forth a new view of science in which the goal is not to seek ultimate explanations but to find the best theoretical account we can of the phenomena of experience and experiment. Newton's more limited goals were not entirely new. They have roots in an earlier scientific tradition that had abandoned the search for the "first springs of natural motions" and other natural phenomena, keeping to the more modest effort to develop the best theoretical account we can: what Richard Popkin calls the "constructive skepticism ... formulated ... in detail by Mersenne and Gassendi," later in Hume's "mitigated skepticism." In this conception, Popkin continues, science proceeds by "doubting our abilities to find grounds for our knowledge, while accepting and increasing the knowledge itself" and recognizing that "the secrets of nature, of things-in-themselves, are forever hidden from us"—the "science without metaphysics ... which was to have a great history in more recent times."[16]

As the impact of Newton's discoveries was slowly absorbed, such lowering of the goals of scientific inquiry became routine. Scientists abandoned the animating idea of the early scientific revolution: that the world will be intelligible to us. It is enough to construct intelligible explanatory theories, a radical difference. By the time we reach Russell's *Analysis of Matter*, he dismisses the very idea of an intelligible world as "absurd," and repeatedly places the word "intelligible" in quotes to highlight the absurdity of the quest. Qualms about action at a distance were "little more than a prejudice," he writes. "If all the world consisted of billiard balls, it would be what is called 'intelligible'—that is, it would never surprise us sufficiently to make us realize that we do not understand it."[17] But even without external surprise, we should recognize how little we understand the world, and should also realize that it does not matter whether we can conceive of how the world works. In his classic introduction to quantum mechanics a few years later, Paul Dirac wrote that physical science no longer seeks to provide pictures of how the world works, that is "a model functioning on essentially classical lines," but only seeks to provide a "way of looking at the fundamental laws which makes their self-consistency obvious." He was referring to the inconceivable conclusions of quantum physics, but could just as readily have said that even the classical Newtonian models had abandoned the hope of rendering natural phenomena intelligible, the primary goal of the early modern scientific revolution, with its roots in common-sense understanding.[18]

---

[16] Joseph Glanvill; see John Henry, "Occult Qualities and the Experimental Philosophy," *History of Science*, xxiv (1986): 335–81. Popkin, *The History of Scepticism from Erasmus to Spinoza* (Berkeley: California UP, 1979), pp. 139f., 213.

It is useful to recognize how radical a shift it was to abandon the mechanical philosophy, and with it any scientific relevance of our common-sense beliefs and conceptions, except as a starting point and spur for inquiry. Galileo scholar Peter Machamer observes that by adopting the mechanical philosophy, and initiating the modern scientific revolution, Galileo had "forged a new model of intelligibility for human understanding, [with] new criteria for coherent explanations of natural phenomena" based on the conception of the world as an elaborate machine. For Galileo, and leading figures in the early modern scientific revolution generally, true understanding requires a mechanical model, a device that an artisan could construct, hence intelligible to us. Thus Galileo rejected traditional theories of tides because we cannot "duplicate [them] by means of appropriate artificial devices."[19]

The model of intelligibility that reigned from Galileo through Newton and beyond has a corollary: when mechanism fails, understanding fails. The apparent inadequacies of mechanical explanation for cohesion, attraction, and other phenomena led Galileo finally to reject "the vain presumption of understanding everything." Worse yet, "there is not a single effect in nature ... such that the most ingenious theorist can arrive at a complete understanding of it."[20] Galileo was formulating a very strong version of what Daniel Stoljar calls "the ignorance hypothesis" in his careful inquiry into the contemporary study of philosophical problems relating to consciousness, concluding that their origins are epistemic and that they are effectively overcome by invoking the ignorance hypothesis—which for Galileo, Newton, Locke, Hume, and others was more than a hypothesis, and extended far beyond the problem of consciousness, encompassing the truths of nature quite generally.[21]

---

[17] Russell, *Analysis of Matter* (London: Allen and Unwin, 1927; reprint, New York: Dover, 1954), pp. 18f., 162.

[18] Dirac, *Principles of Quantum Mechanics* (New York: Oxford, 1930), p. 10, brought to my attention by John Frampton.

[19] Machamer, Introduction, and "Galileo's Machines, His Mathematics, and His Experiments," in Machamer, ed., *The Cambridge Companion to Galileo* (New York: Cambridge, 1998), pp. 17, 69.

[20] *Ibid.*; cited by Pietro Redondi, "From Galileo to Augustine."

Though much more optimistic than Galileo about the prospects for mechanical explanation, Descartes too recognized the limits of our cognitive reach. Rule 8 of the *Regulae* reads: "If in the series of subjects to be examined we come to a subject of which our intellect cannot gain a good enough intuition, we must stop there; and we must not examine the other matters that follow, but must refrain from futile toil." Specifically, Descartes speculated that the workings of *res cogitans* may lie beyond human understanding. He thought that we may not "have intelligence enough" to understand the workings of mind, in particular, the normal use of language, with its creative aspects, his core example: the capacity of every human, but no beast-machine, to use language in ways appropriate to situations but not caused by them, and to formulate and express coherent thoughts without bound, perhaps "incited or inclined" to speak in certain ways by internal and external circumstances but not "compelled" to do so, as his followers put the matter.[22]

However, Descartes continued, even if the explanation of normal use of language and other forms of free and coherent choice of action lies beyond our cognitive grasp, that is no reason to question the authenticity of our experience. Quite generally, "free will" is "the noblest thing" we have, Descartes held: "there is nothing we comprehend more evidently and more perfectly" and "it would be absurd" to doubt something that "we comprehend intimately, and experience within ourselves" (that "the free actions of men [are] undetermined") merely because it conflicts with something else "which we know must be by its nature incomprehensible to us" ("divine preordination").[23]

---

[21] Stoljar, *Ignorance and Imagination* (New York: Oxford, 2006). Recall that Newton hoped that there might be a scientific (that is, mechanical) solution to the problems of matter and motion.

[22] On these topics, see my *Cartesian Linguistics* (New York: Harper and Row, 1966); third edition, edited by James McGilvray, with introduction, full translations and quotes from updated scholarly editions (Cybereditions, 2008); and *Language and Mind* (New York: Harcourt Brace Jovanovich, 1968), chapter 1. Note that the concerns go far beyond indeterminacy of free action, as is particularly evident in the experimental programs by Cordemoy and others on "other minds"; see *Cartesian Linguistics*.

[23] Descartes, Letter to Queen Christina of Sweden, 1647; *Principles of Philosophy*. For discussion, see Tad Schmaltz, *Malebranche's Theory of the Soul* (New York: Oxford, 1996), pp. 204ff.

Such thoughts about cognitive limits do not comport well with Descartes's occasional observation that human reason "is a universal instrument which can serve for all contingencies," whereas the organs of an animal or machine "have need of some special adaptation for any particular action." But let us put that aside, and keep to the more reasonable conclusions about cognitive limits.

The creative use of language was a basis for what has been called the "epistemological argument" for mind-body dualism and also for the scientific inquiries of the Cartesians into the problem of "other minds"—much more sensible, I believe, than contemporary analogues, often based on misinterpretation of a famous paper of Turing's, a topic that I will put aside.[24]

Desmond Clarke is accurate, I think, in concluding that "Descartes identified the use of language as the critical property that distinguishes human beings from other members of the animal kingdom and [that] he developed this argument in support of the real distinction of mind and matter." I think he is also persuasive in interpreting the general Cartesian project as primarily "natural philosophy" (science), an attempt to press mechanical explanation to its limits; and in regarding the *Meditations* "not as the authoritative expression of Descartes's philosophy, but as an unsuccessful attempt to reconcile his theologically suspect natural philosophy with an orthodox expression of scholastic metaphysics."[25] In pursuing his natural science, Descartes tried to show that mechanical explanation reached very far, but came to an impassable barrier in the face of such mental phenomena as the creative use of language. He therefore, quite properly, adopted the standard scientific procedure of seeking some new principles to account for such mental phenomena—a quest that lost one primary motivation when mechanical explanation was demonstrated to fail for everything.

Clarke argues that "Descartes's dualism was an expression of the extent of the theoretical gap between [Cartesian physics] and the descriptions of mental life that we formulate from the first person perspective of our own thinking." The gap therefore results from Descartes's "impoverished concept of matter" and can be overcome by "including new theoretical entities in one's concept of matter."[26] Whether the latter speculation is correct or not, it does not quite capture the deficiencies of classical science from Galileo through Newton and beyond. The underlying concept of matter and motion—based on conceivability, intelligibility, and common-sense understanding—had to be abandoned, and science had to proceed on an entirely new course in investigating the simplest phenomena of motion, and all other aspects of the world, including mental life.

---

[24] See my "Turing on the 'Imitation Game'," in Stuart Schieber, ed., *The Turing Test* (Cambridge: MIT, 2004), pp. 317–21.

[25] Clarke, *Descartes's Theory of Mind* (New York: Oxford, 2003), p. 12. See also Descartes's 1641 letter to Mersenne on the goal of the *Meditations*, cited by Margaret Wilson, *Descartes* (Boston: Routledge Kegan Paul, 1978), p. 2.

[26] Clarke, *op. cit.*, p. 258.

Despite the centrality of the creative use of language to Cartesian science, it was only one illustration of the general problem of will, and choice of appropriate action, which remains as mysterious to us as it was to seventeenth century scientists, so it seems to me, despite sophisticated arguments to the contrary. The problems are hardly even on the scientific agenda. There has been very valuable work about how an organism executes a plan for integrated motor action—say, how a person reaches for a cup on the table. But no one even raises the question of why this plan is executed rather than some other one, apart from the very simplest organisms and special circumstances of motivation. Much the same is true even for visual perception. Cognitive neuroscientists Nancy Kanwisher and Paul Downing reviewed research on a problem posed in 1850 by Helmholtz: "even without moving our eyes, we can focus our attention on different objects at will, resulting in very different perceptual experiences of the same visual field." The phrase "at will" points to an area beyond serious empirical inquiry, still the mystery it was for Newton at the end of his life when he continued to seek some "subtle spirit" that lies hidden in all bodies and that might, without "absurdity," account for their properties of attraction and repulsion, along with the nature and effects of light, sensation, and the way "members of animal bodies move at the command of the will"—all comparable mysteries for Newton, perhaps even beyond our understanding.[27]

It has become standard practice in recent years to describe the problem of consciousness as "the hard problem," others being within our grasp, now or down the road. I think there are reasons for some skepticism; particularly when we recognize how sharply understanding declines beyond the simplest systems of nature. To illustrate with a few examples, a review-article by Eric Kandel and Larry Squire on the current state of efforts aimed at "breaking down scientific barriers to the study of brain and mind" concludes that "the neuroscience of higher cognitive processes is only beginning."[28] C.R. Gallistel points out that "we clearly do not understand how the nervous system computes," or even "the foundations of its ability to compute," even for "the small set of arithmetic and logical operations that are fundamental to any computation."

---

[27] Kanwisher and Downing, "Separating the Wheat from the Chaff," *Science*, CCLXXXII (October 2, 1998). Newton, General Scholium.

[28] Kandel and Squire, "Neuroscience," *Science*, CCXC (November 10, 2000): 1113–20.

Reviewing the remarkable computational capacities of insects, he concludes that it is a mistake to suppose that the nervous system does not carry out complex symbolic computations on grounds of "our inability, as yet to understand how the nervous system computes at the cellular and molecular level.... We do not know what processes belong to the basic instruction set of the nervous system—the modest number of elementary operations built into the hardware of any computing device...."[29] Semir Zeki, who is optimistic about the prospects for bringing the brain sciences to bear even on creativity in the visual arts, nevertheless reminds us that "how the brain combines the responses of specialized cells to indicate a continuous vertical line is a mystery that neurology has not yet solved," or even how one line is differentiated from others or from the visual surround. Basic traditional questions are not even on the research agenda, and even simple ones that might be within reach remain baffling.[30]

It is common to assert that "the mental is the neurophysiological at a higher level." To entertain the idea makes sense, but for the present, only as a guide to inquiry, without much confidence about what "the neurophysiological" will prove to be. Similarly it is premature to hold that "it is empirically evident that states of consciousness are the necessary consequence of neuronal activity." Too little is understood about the functioning of the brain.[31]

History also suggests caution. In early modern science, the nature of motion was the "hard problem." "Springing or Elastic Motions" is the "hard rock in Philosophy," Sir William Petty observed, proposing ideas resembling those soon developed much more richly by Newton. The "hard problem" was that bodies that seem to our senses to be at rest are in a "violent" state, with "a strong endeavor to fly off or recede from one another," in Robert Boyle's words. The problem, he felt, is as obscure as "the Cause and Nature" of gravity, thus supporting his belief in "an intelligent Author or Disposer of Things." Even the skeptical Newtonian Voltaire argued that the ability of humans to "produce a movement" where there was none shows that "there is a God who gave movement" to matter, and "so far are we from conceiving what matter is" that we do not even know if there is any "solid matter in the universe." Locke relinquished to divine hands "the gravitation of matter towards matter, by ways, inconceivable to me." Kant rephrased the "hard problem," arguing that to reach his conclusions, Newton was compelled to tacitly "assume that all matter exercises this motive force [of universal attraction] simply as matter and by its essential nature"; by rejecting the assumption, he was "at variance with himself," caught in a contradiction. Newton therefore did not, as he claimed, really leave "the physicists full freedom to explain the possibility of such attraction as they might find good, without mixing up his propositions with their play of hypotheses." Rather, "The concept of matter is reduced to nothing but moving forces.... The attraction essential to all matter is an immediate action of one matter on another across empty space," a notion that would have been anathema to the great figures of seventeenth century science, "such Masters, as the Great Huygenius, and the incomparable Mr. Newton," in Locke's words.[32]

---

[29] Gallistel, "Neurons and Memory," in M. Gazzaniga, ed., *Conversations in the Cognitive Neurosciences* (Cambridge: MIT, 1997), pp. 71–89; "Symbolic Processes in the Insect Brain," in D. Scarborough and S. Sternberg, eds., *Methods, Models and Conceptual Issues: An Invitation to Cognitive Science*, Volume 4 (Cambridge: MIT, 1998), pp. 1–51.

[30] Zeki, "Art and the Brain," *Daedalus: The Brain* (Spring 1998): 71–104.

[31] Nagel, *op. cit.* For some cautionary notes on "sharp logical separation between the nervous system and the rest of the organism," see Charles Rockland, *The Nematode as a Model Complex System*, Working Paper (LIDS-WP-1865), Laboratory for Information and Decisions Systems, MIT (April 14, 1989), p. 30.

The "hard problems" of the day were not solved; rather abandoned, as, over time, science turned to its more modest post-Newtonian course. Friedrich Lange, in his classic nineteenth century history of materialism, observed that we have "so accustomed ourselves to the abstract notion of forces, or rather to a notion hovering in a mystic obscurity between abstraction and concrete comprehension, that we no longer find any difficulty in making one particle of matter act upon another without immediate contact, ... through void space without any material link. From such ideas the great mathematicians and physicists of the seventeenth century were far removed. They were all in so far genuine Materialists in the sense of ancient Materialism that they made immediate contact a condition of influence." This transition over time is "one of the most important turning-points in the whole history of Materialism," depriving the doctrine of much significance, if any at all. Newton not only joined the great scientists of his day in regarding "the now prevailing theory of *actio in distans*... simply as absurd, [but] also felt himself obliged, in the year 1717, in the preface to the second edition of his 'Optics,' to protest expressly against [the] view" of his followers who "went so far as to declare gravity to be a fundamental force of matter," requiring no "further mechanical explanation from the collision of imponderable particles." Lange concludes that "the course of history has eliminated this unknown material cause [that so troubled Newton], and has placed the mathematical law itself in the rank of physical causes." Hence "What Newton held to be so great an absurdity that no philosophic thinker could light upon it, is prized by posterity as Newton's great discovery of the harmony of the universe!"[33] The conclusions are commonplace in the history of science. Fifty years ago Alexander Koyré observed that despite his unwillingness to accept the conclusion, Newton had demonstrated that "a purely materialistic pattern of nature is utterly impossible (and a purely materialistic or mechanistic physics, such as that of Lucretius or of Descartes, is utterly impossible, too)"; his mathematical physics required the "admission into the body of science of incomprehensible and inexplicable 'facts' imposed up on us by empiricism," by what is observed and our conclusions from these observations.[34]

---

[32] Henry, *op. cit.* Alan Kors, "The Atheism of D'Holbach and Naigeon," in M. Hunger and D. Wootton, eds., *Atheism from the Reformation to the Enlightenment* (New York: Oxford, 1992), pp. 273–300. Locke, see above, and Yolton, *op. cit.*, p. 199. Voltaire, Kant, McMullin, *op. cit.*, pp. 113, 122f., from Kant's *Metaphysical Foundations of Natural Science*, 1786; Michael Friedman, "Kant and Newton: Why Gravity Is Essential to Matter," in P. Bricker and R.I.G. Hughes, eds., *Philosophical Perspectives on Newtonian Science* (Cambridge: MIT, 1990), pp. 185–202. Howard Stein, "On Locke, 'the Great Huygenius, and the incomparable Mr. Newton'," in *ibid.* Friedman argues that there is no contradiction between Newton and Kant because they do not mean the same thing by "essential," Kant having discarded Newton's metaphysics and making an epistemological point within his "Copernican revolution in metaphysics."

George Coyne describes it as "paradoxical that the rise of materialism as a philosophy in the seventeenth and eighteenth centuries is attributed to the birth of modern science, when in reality matter as a workable concept had been eliminated from scientific discourse" with the collapse of the mechanical philosophy.[35] Also paradoxical is the influence of Gilbert Ryle's ridicule of the "ghost in the machine," quite apart from the accuracy of his rendition of the Cartesian concepts. It was the machine that Newton exorcised, leaving the ghost intact. The "hard problem" of the materialists disappeared, and there has been little noticeable progress in addressing other "hard problems" that seemed no less mysterious to Descartes, Newton, Locke, and other leading figures.

---

[33] Lange, *Geschichte des Materialismus und Kritik seiner Bedeutung in der Gegenwart* (1865); *The History of Materialism* (London: Kegan Paul, Trench, Trubner, 1925, translation of third expanded edition).

[34] Koyré, *From the Closed World to the Infinite Universe* (Baltimore: Johns Hopkins, 1958), p. 210.

[35] Coyne, "The Scientific Venture and Materialism: False Premises," in *Space or Spaces as Paradigms of Mental Categories* (Milan: Fondazione Carlo Erba, 2000), pp. 7–19.

The third English edition of Lange's much expanded history of materialism appeared in 1925 with an introduction by Bertrand Russell, who shortly after published his *Analysis of Matter.* Developing his neutral monism, Russell carried further seventeenth and eighteenth century skepticism about matter, and recognition of the plausibility (or for some necessity) of thinking matter. Russell held that there are "three grades of certainty. The highest grade belongs to my own percepts; the second grade to the percepts of other people; the third to events which are not percepts of anybody," constructions of the mind established in the course of efforts to make sense of what we perceive. "A piece of matter is a logical structure composed of [such] events," he therefore concluded. We know nothing of the "intrinsic character" of such mentally constructed entities, so there is "no ground for the view that percepts cannot be physical events." For science to be informative, it cannot be restricted to structural knowledge of such logical properties. Rather, "the world of physics [that we construct] must be, in some sense, continuous with the world of our perceptions, since it is the latter which supplies the evidence for the laws of physics." The percepts that are required for this task—perhaps just meter-readings, Eddington had argued shortly before—"are not known to have any intrinsic character which physical events cannot have, since we do not know of any intrinsic character which could be incompatible with the logical properties that physics assigns to physical events." Accordingly, "What are called 'mental' events...are part of the material of the physical world." Physics itself seeks only to discover "the causal skeleton of the world, [while studying] percepts only in their cognitive aspect; their other aspects lie outside its purview"—though we recognize their existence, at the highest grade of certainty in fact.[36]

The basic conundrum recalls a classical dialogue between the intellect and the senses, in which the intellect says that color, sweetness, and the like are only convention while in reality there are only atoms and the void, and the senses reply: "Wretched mind, from us you are taking the evidence by which you would overthrow us? Your victory is your own fall."[37]

To illustrate his conclusion, Russell asks us to consider a blind physicist who knows the whole of physics but does not have "the knowledge which [sighted] men have" about, say, the quality of the color blue. In their review of related issues, Daniel Stoljar and Yujin Nagasawa call this the "knowledge intuition," as distinct from the "knowledge argument," presented in the resurrection of Russell's example by Frank Jackson: in this case the physicist (Mary) "learns everything there is to know about the physical nature of the world" while confined to a black-and-white room, but when released "will learn what it is like to see something red."[38]

---

[36] Russell, *Analysis of Matter,* chapter 37. Russell did not work out how percepts in their cognitive aspect were assimilated into the "causal skeleton of the world," leaving him open to a counterargument by mathematician Max Newman. See Russell's letter to Newman, 24 April 1928, in *The Autobiography of Bertrand Russell: 1914–1944,* Volume II (Boston: Little Brown, 1967).

[37] Democritus, quoted by Erwin Schrödinger, *Nature and the Greeks* (New York: Cambridge, 1954), p. 89. Brought to my attention by Jean Bricmont.

There is a substantial literature seeking to evade the argument. One popular though contested proposal is that what Mary lacks is not the knowledge of the world that we have, but a range of abilities, a species of "knowing how." That seems unhelpful, in part because there is an irreducible cognitive element in "knowing how," which goes beyond abilities; but also for the kinds of reasons that Hume discussed in connection with moral judgments. Since these, he observed, are unbounded in scope and applicable to new situations, they must be based on a finite array of general principles (which are, furthermore, part of our nature though they are beyond the "original instincts" shared with animals). The knowledge that *we* have but Mary lacks is a body of knowledge that does not fall within the knowing-how/knowing-that dichotomy: it is knowledge *of*—knowledge of rules and principles that yield unbounded capacities to act appropriately. All of this is for the most part unconscious and inaccessible to consciousness, as in the case of knowledge of the rules of language, vision, and so on. Such conclusions have been rejected as a matter of principle by Quine, Searle, and many others, but not convincingly or even coherently, I think.[39]

Russell's knowledge intuition led him to conclude that physics has limits: experience in general lies "outside its purview" apart from cognitive aspects that provide empirical evidence, though along with other mental events, experience is "part of the material of the physical world," a phrase that seems to mean no more than "part of the world." We must have "an interpretation of physics which gives a due place to perceptions," Russell held, or it has no empirical basis. Jackson's knowledge argument leads him to the conclusion that "physicalism is false." Or in a later version, that to be valid "materialism [as] a metaphysical doctrine" must incorporate "the psychological story about our world"; the "story about our world told purely in physical terms [must] enable one to deduce the phenomenal nature of psychological states."[40] But that is uninformative until some clear concept of physicalism/materialism is offered. Classical interpretations having vanished, the notions of body, material, physical are hardly more than honorific designations for what is more or less understood at some particular moment in time, with flexible boundaries and no guarantee that there will not be radical revision ahead, even at its core. If so, the knowledge argument only shows (with Russell) that humanly-constructed physics has limits, or that Mary did not know all of physics (she had not drawn the right conclusions from Eddington's meter readings).

---

[38] Stoljar and Nagasawa, "Introduction," in Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, *There's Something about Mary* (Cambridge: MIT, 2004), pp. 1–36.

[39] Hume, see John Mikhail, *Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in A Theory of Justice* (Ph.D. Dissertation, Cornell University, 2000); *Moral Grammar: Rawls' Linguistic Analogy and the Cognitive Science of Moral Judgment* (New York: Cambridge, in press); "Universal Moral Grammar: Theory, Evidence, and the Future," *Trends in Cognitive Sciences*, April 2007. On the irrelevance (and as it is formulated, even incoherence) of the doctrine of "accessibility to consciousness"), see my *Reflections on Language* (New York: Pantheon, 1975), *Rules and Representations* (New York: Columbia, 1980), *New Horizons in the Study of Language and Mind* (New York: Cambridge, 2000). On the rules of visual perception, inaccessible to consciousness in the interesting cases, see Donald Hoffman, *Visual Intelligence* (New York: Norton, 1998).

To resurrect something that resembles a "mind-body problem," it would be necessary to characterize *physicalism* (*matter*, and so on) in some post-Newtonian fashion, or to argue that the problem arises even if the concepts are abandoned. Both approaches have been pursued. I will return to current examples. An alternative approach is to dismiss the mind-body problem, and to approach the knowledge intuition/argument as a problem of the natural sciences. Rephrasing Russell's thought experiment, we might say that like all animals, we have internal capacities that reflexively provide us with what ethologists called an *Umwelt*, a world of experience, different for us and for bees—in fact, differing among humans, depending on what they understand. That is why radiology is a medical specialty. Galileo saw the moons of Jupiter through his primitive telescope, but those he sought to convince could only see magnification of terrestrial objects, and took his telescope to be a conjuring trick (at least if Paul Feyerabend's reconstruction of the history is correct). What I hear as noise is perceived as music by my teenage grandchildren, at a fairly primitive level of perceptual experience. And so on quite generally.

Being reflective creatures, unlike others, we go on to seek to gain a deeper understanding of the phenomena of experience. These exercises are called myth, or magic, or philosophy, or science. They reveal not only that the world of experience is itself highly intricate and variable, resulting from the interaction of many factors, but also that the modes of interpretation that intuitive common sense provides do not withstand analysis, so that the goals of science must be lowered in the manner recognized in post-Newtonian science. From this point of view, there is no objective science from a third-person perspective, just various first-person perspectives, matching closely enough among humans so that a large range of agreement can be reached, with diligence and cooperative inquiry. Being inquisitive as well as reflective creatures, if we can construct a degree of theoretical understanding in some domain, we try to unify it with other branches of inquiry, reduction being one possibility but not the only one.

---

[40] Jackson, "What Mary Didn't Know," "Postscript," in Ludlow et al., *op. cit.*, pp. xv–xix and 410–42.

We can anticipate that our quest might fail, for one reason, because our basically shared capacities of understanding and explanation have limits—a truism that is sometimes thoughtlessly derided as "mysterianism," though not by Descartes and Hume, among others. It could be that these innate cognitive capacities do not lead us beyond some understanding of Russell's causal skeleton of the world (and enough about perception to incorporate evidence within this mental construction), and it is an open question how much of that can be attained. In principle, the limits could become topics of empirical inquiry into the nature of what we might call "the science-forming faculty," another "mental organ." These are interesting topics, but the issues are distinct from the traditional mind-body problem, which evaporated after Newton, or from the question of how mental aspects of the world, including direct experience, relate to the brain, one of the many problems of unification that arise in the sciences.

In brief, if we are biological organisms, not angels, much of what we seek to understand might lie beyond our cognitive limits—maybe a true understanding of anything, as Galileo concluded, and Newton in a certain sense demonstrated. That cognitive reach has limits is not only a truism, but also a fortunate one: if there were no limits to human intelligence, it would lack internal structure, and would therefore have no scope: we could achieve nothing by inquiry. The basic points were expressed clearly by Charles Sanders Peirce in his discussion of the need for innate endowment that "puts a limit upon admissible hypotheses" if knowledge is to be acquired.[41] Similarly if a zygote had no further genetic instructions constraining its developmental path, it would at best grow into a creature formed solely by physical law, like a snowflake, nothing viable.

---

[41] Peirce, "The Logic of Abduction," in V. Tomas, ed., *Peirce's Essays in the Philosophy of Science* (New York: Liberal Arts Press, 1957), pp. 235–55. See my *Language and Mind*, pp. 90ff. for discussion of his proposals, and fallacies invoking natural selection that lead him to the ungrounded (and implausible) belief that our "guessing instinct" leads us to true theories.

We might think of the natural sciences as a kind of chance convergence between our cognitive capacities and what is more or less true of the natural world. There is no reason to believe that humans can solve every problem they pose or even that they can formulate the right questions; they may simply lack the conceptual tools, just as rats cannot deal with a prime number maze.

Russell's general conclusions seem to me on the right track. The formulation can be improved, I think, by simply dropping the words "matter" and "physical." Since the Newtonian revolution, we speak of the "physical" world much as we speak of the "real" truth: for emphasis, but adding nothing. We can distinguish various aspects of the world—say chemical, electrical, experiential, and the rest—and we can then inquire into their underlying principles and their relations with other systems, problems of unification.

Suppose we adopt the "mitigated skepticism" that was warranted after Newton, if not before. For the theory of mind, that means following Gassendi's advice in his *Objections*. He argued that Descartes had at most shown "the perception of the existence of mind, [but] fail[ed] to reveal its nature." It is necessary to proceed as we would in seeking to discover "a conception of Wine superior to the vulgar," by investigating how it is constituted and the laws that determine its functioning. Similarly, he urged Descartes, "it is incumbent on you, to examine yourself by a certain chemicallike labor, so that you can determine and demonstrate to us your internal substance"[42]—and that of others.

The theory of mind can be pursued in many ways, like other branches of science, with an eye to eventual unification, whatever form it may take, if any. That is the task that Hume undertook when he investigated what he called "the science of human nature," seeking "the secret springs and principles, by which the human mind is actuated in its operations," including those "parts of [our] knowledge" that are derived from "the original hand of nature," an enterprise he compared to Newton's; essentially what in contemporary literature is termed "naturalization of philosophy" or "epistemology naturalized." Gassendi's recommended course was in fact being pursued in the "cognitive revolution" of the seventeenth century by British neoplatonists and continental philosophers of language and mind, and has been taken up with renewed vigor in recent years, but I will put that matter aside.[43]

---

[42] Cited by Wilson, *op. cit.*, p. 95.

Chemistry itself quite explicitly pursued this course. The eighteenth century chemist Joseph Black recommended that "chemical affinity be received as a first principle, which we cannot explain any more than Newton could explain gravitation, and let us defer accounting for the laws of affinity, till we have established such a body of doctrine as he has established concerning the laws of gravitation." Being yet "very far from the knowledge of first principles," chemical science should be "analytical, like Newton's *Optics*, in the form of a general law, at the very end of our induction, as the reward of our labour." The course he outlined is the one that was actually followed, as chemistry established a rich body of doctrine, its "triumphs...built on no reductionist foundation but rather achieved in isolation from the newly emerging science of physics," historian of chemistry Arnold Thackray observes. Newton and his followers did attempt to "pursue the thoroughly Newtonian and reductionist task of uncovering the general mathematical laws which govern all chemical behavior" and to develop a principled science of chemical mechanisms based on physics and its concepts of interactions among "the ultimate permanent particles of matter." But the Newtonian program was undercut by Dalton's "astonishingly successful weight-quantification of chemical units," Thackray continues, shifting "the whole area of philosophical debate among chemists from that of chemical *mechanisms* (the *why?* of reaction) to that of chemical *units* (the *what?* and *how much?*)," a theory that "was profoundly antiphysicalist and anti-Newtonian in its rejection of the unity of matter, and its dismissal of short-range forces." "Dalton's ideas were chemically successful. Hence they have enjoyed the homage of history, unlike the philosophically more coherent, if less successful, reductionist schemes of the Newtonians."[44]

Adopting contemporary terminology, we might say that Dalton disregarded the explanatory gap between chemistry and physics by ignoring the underlying physics, much as post-Newtonian physicists disregarded the explanatory gap between Newtonian dynamics and the mechanical philosophy by ignoring (and in this case rejecting) the latter, though it was self-evident to common-sense understanding.

---

[43] Hume, *Inquiry*, II.1. On dubious modern efforts to formulate what had been a reasonably clear project before the separation of philosophy from science, see *New Horizons*, pp. 79f., 144f.; and generally chapters 5, 6 (reprinted from *Mind*, CIV (1995): 1–61).

[44] Black, Robert Schofield, *Mechanism and Materialism* (Princeton: University Press, 1970), p. 226; William Brock, *The Norton History of Chemistry* (New York: Norton, 1993), p. 271. Thackray, *Atoms and Powers* (Cambridge: Harvard, 1970), pp. 37f., 276f.

That has often been the course of science since, though not without controversy and sharp criticism, often later recognized to have been seriously misguided. Well into the twentieth century the failure of reduction of chemistry to physics was interpreted by prominent scientists as a critically important explanatory gap, showing that chemistry provides "merely classificatory symbols that summarized the observed course of a reaction," to quote Brock's standard history. Kekulé, whose structural chemistry was an important step towards eventual unification of chemistry and physics, doubted that "absolute constitutions of organic molecules could ever be given"; his models and analysis of valency were to have an instrumental interpretation only, as calculating devices. Lavoisier before him believed that "the number and nature of elements [is] an unsolvable problem, capable of an infinity of solutions none of which probably accord with Nature"; "It seems extremely probable that we know nothing at all about...[the]... indivisible atoms of which matter is composed," and never will, he believed. Kekulé seems to be saying that there is not a problem to be solved; the structural formulas are useful or not, but there is no truth of the matter. Large parts of physics were understood the same way. Poincaré went so far as to say that we adopt the molecular theory of gases only because we are familiar with the game of billiards. Boltzmann's scientific biographer speculates that he committed suicide because of his failure to convince the scientific community to regard his theoretical account of these matters as more than a calculating system—ironically, shortly after Einstein's work on Brownian motion and broader issues had convinced physicists of the reality of the entities he postulated. Bohr's model of the atom was also regarded as lacking "physical reality" by eminent scientists. In the 1920s, America's first Nobel Prize-winning chemist dismissed talk about the real nature of chemical bonds as metaphysical "twaddle": they are nothing more than "a very crude method of representing certain known facts about chemical reactions, a mode of representation" only, because the concept could not be reduced to physics. The rejection of that skepticism by a few leading scientists, whose views were condemned at the time as a conceptual absurdity, paved the way for the eventual unification of chemistry and physics, with Linus Pauling's quantum-theoretic account of the chemical bond seventy years ago.[45]

---

[45] Brock, *op. cit.* For sources and further discussion, see *New Horizons*; my *Knowledge of Language* (New York: Praeger, 1986), pp. 251–52. David Lindley, *Boltzmann's Atom* (New York: Free Press, 2001). Some argue that even if quantum-theoretic unification succeeds, "in some sense the program of reduction of chemistry to [the new] physics fails," in part because of "practical issues of intractability." See Maureen Christie and John Christie, "'Laws' and 'Theories' in Chemistry Do Not Obey the Rules," in Nalin Bhushan and Stuart Rosenfield, eds., *Of Minds and Molecules* (New York: Oxford, 2000), pp. 34–50.

In 1927, Russell observed that chemical laws "cannot at present be reduced to physical laws,"[46] an observation that was found to be misleading: the words "at present" turned out to understate the matter. Chemical laws could not ever be reduced to physical laws, because the conception of physical laws was erroneous. The perceived explanatory gap was never filled. It was necessary, once again, to dismiss as irrelevant the notion of "conceivability" and "intelligibility of the world," in favor of the mitigated skepticism of methodological naturalism: seeking to increase our knowledge while keeping an open mind about the possibility of reduction.

There are fairly clear parallels to contemporary discussion of language and mind, and some lessons that can be drawn. The study of insect symbolic representation, organization of motor behavior, mammalian vision, human language, moral judgment, and other topics is in each case well advised to follow Joseph Black's prescription. If these inquiries succeed in developing a "body of doctrine" that accounts for elements of insect navigation, or the rule that image motions are interpreted (if other rules permit) as rigid motions in three dimensions, or that displacement operations in language observe locality principles, and so on, that should be regarded as normal science, even if unification with neurophysiology has not been achieved—and might not be for a variety of possible reasons, among them that the expected "reduction base" is misconceived and has to be modified. Needless to say, the brain sciences are not as firmly established as basic physics was a century ago, or as the mechanical philosophy was in Newton's day. It is also pointless to insist on doctrines about accessibility to consciousness: even if they could be given a coherent formulation, they would have no bearing on the "physical reality" of the rigidity principle or locality conditions. We should understand enough by now to dismiss the interpretation of theoretical accounts as no more than a way of "representing certain known facts about [behavior], a mode of representation" only—a critique commonly leveled against theories of higher mental faculties, though not insect computation, another illustration of the methodological dualism that is so prevalent in critical discussion of inquiry into language and mind.[47]

---

[46] Russell, *Analysis of Matter*, p. 388.

[47] See references of note 39. Sometimes misunderstanding and distortion reach the level of the surreal. For some startling examples, see my contribution to the Symposium on Margaret Boden, *Mind as Machine: A History of Cognitive Science* (New York: Oxford, 2006), in *Artificial Intelligence*, CLXXI (2007): 1094–103 (Elsevier), at www.sciencedirect.com. On "the rigidity rule and [Shimon] Ullman's theorem," see Hoffman, *op. cit.*, p. 159. Needless to say, the rule is inaccessible to consciousness.

It is also instructive to observe the reemergence of much earlier insights, though divorced from their grounding in the collapse of traditional physicalism. Thus we read today of the thesis of the new biology that "Things mental, indeed minds, are emergent properties of brains, [though] these emergences are ... produced by principles that ... we do not yet understand"—neuroscientist Vernon Mountcastle, formulating the guiding theme of a collection of essays reviewing the results of the Decade of the Brain that ended the twentieth century. The phrase "we do not *yet* understand" might well suffer the same fate as Russell's similar comment about chemistry seventy years earlier. Many other prominent scientists and philosophers have presented essentially the same thesis as an "astonishing hypothesis" of the new biology, a "radical" new idea in the philosophy of mind, "the bold assertion that mental phenomena are entirely natural and caused by the neurophysiological activities of the brain," opening the door to novel and promising inquiries, a rejection of Cartesian mind-body dualism, and so on.[48] In fact, all reiterate, in virtually the same words, formulations of centuries ago, after the traditional mind-body problem became unformulable with the disappearance of the only coherent notion of body (physical, material, and so on): for example Joseph Priestley's conclusion that properties "termed mental" reduce somehow to "the organical structure of the brain,"[49] stated in different words by Hume, Darwin, and many others, and almost inescapable, it would seem, after the collapse of the mechanical philosophy.

Priestley's important work was the culmination of a century of reflections on Locke's speculation, and their most elaborate development.[50] He made it clear that his conclusions about thinking matter followed directly from the collapse of any serious notion of *body*, or *matter*, or *physical*. He wrote that "The principles of the Newtonian philosophy were no sooner known, than it was seen how few in comparison, of the phenomena of Nature were owing to solid matter, and how much to powers which were only supposed to accompany and surround the solid parts of matter .... Now when solidity had apparently so very little to do in the system, it is really a wonder that it did not occur to philosophers sooner ... that there might be no such thing in Nature."

---

[48] Mountcastle, *Daedalus: The Brain* (Spring 1988). For sources, see *New Horizons*, chapter 5.

[49] Priestley, *Materialism*, from *Disquisitions Relating to Matter and Spirit*, 1777. In John Passmore, ed., *Priestley's Writings on Philosophy, Science, and Politics* (New York: Collier-MacMillan, 1965).

[50] Similar ideas appear pre-Newton, particularly in the Objections to the *Meditations*, where critics ask how Descartes can know, "without divine revelation...that God has not implanted in certain bodies a power or property enabling them to doubt, think, etc." See Catherine Wilson, in Strawson, *op. cit.*

There is, then, no longer any reason to suppose that "the principle of thought or sensation [is] incompatible with matter," Priestley concluded. Accordingly, "the whole argument for an immaterial thinking principle in man, on this supposition, falls to the ground; matter, destitute of what has hitherto been called solidity, being no more incompatible with sensation and thought than that substance which without knowing anything farther about it, we have been used to call immaterial." The powers of sensation, perception, and thought reside in "a certain organized system of matter, [and] necessarily exist in, and depend upon, such a system." It is true that "we have a very imperfect idea of what the power of perception is," and that we may never attain a "clear idea," but "this very ignorance ought to make us cautious in asserting with what other properties it may, or may not, exist." Only a "precise and definite knowledge of the nature of perception and thought can authorize any person to affirm whether they may not belong to an extended substance which also has the properties of attraction and repulsion." Our ignorance provides no warrant for supposing that sensation and thought are incompatible with post-Newtonian matter. "In fact, there is the same reason to conclude, that the powers of sensation and thought are the necessary result of a particular organization, as that sound is the necessary result of a particular concussion of the air." And in a later discussion, "In my opinion there is just the same reason to conclude that the brain *thinks*, as that it is *white*, and *soft*."[51]

Priestley criticizes Locke for being hesitant in putting forth his speculation about thinking matter, since the conclusion follows so directly from "the universally accepted rules of philosophizing such as are laid down by Sir Isaac Newton." He urges that we abandon the methodological dualism that deters us from applying to thought and sensation the rules that we follow "in our inquiries into the causes of particular appearances in nature," and expresses his hope "that when this is plainly pointed out the inconsistency of our conduct cannot fail to strike us and be the means of inducing" philosophers to apply the same maxim to investigation of mental aspects of the world that they do in other domains—a hope that has yet to be realized, I think.[52]

---

[51] Priestley, *Materialism.* For later discussion, see Yolton, *Thinking Matter*, p. 113. Similar conclusions had been drawn by La Mettrie a generation earlier, but in a different framework and without addressing the Cartesian arguments to which he is attempting to respond. The same is true of Gilbert Ryle and other modern attempts. For some discussion, see *Cartesian Linguistics.*

[52] For discussion and illustrations, see *New Horizons.* See Strawson, below, on "hyperdualism."

Priestley clearly "wished the disappearance of solid matter to signal an end to matter-spirit dualism," Thackray writes. And with it an end to any reason to question the thesis of thinking matter.[53] In John Yolton's words, Priestley's conclusion was "not that all reduces to matter, but rather that the kind of matter on which the two-substance view is based does not exist," and "with the altered concept of matter, the more traditional ways of posing the question of the nature of thought and of its relations to the brain do not fit. We have to think of a complex organized biological system with properties the traditional doctrine would have called mental *and* physical."[54] Priestley's conclusions are essentially those reached by Eddington and Russell, and developed in recent years particularly by Galen Strawson and Daniel Stoljar, in ways to which we return.

Reviewing the development of Locke's suggestion in England through the eighteenth century, Yolton observes that "Priestley's fascinating suggestions were not taken up and extended; they were hardly even perceived as different from earlier versions of materialism. The issues raised by Locke's suggestion of thinking matter...played themselves out through the century, but no one gave the emerging view of man as one substance—foreshadowed by Priestley—a systematic articulation."[55] A conclusion that remains largely true, even for simple organisms, if we interpret it as referring to the unification problem.

Having argued that the mind-body problem disappears when we follow the "principles of the Newtonian philosophy," Priestley turns to confronting efforts to reconstitute something that resembles the problem, even after one of its terms—body (matter, and so on)—no longer has a clear sense. The first is "the difficulty of conceiving how thought can arise from matter ... an argument that derives all its force from our ignorance," he writes, and has no force unless there is a demonstration that they are "absolutely incompatible with one another." Priestley was not troubled by qualms arising from ignorance, rightly I think, any more than scientists should have been concerned about irreducibility of the mysterious properties of matter and motion to the mechanical philosophy, or in more modern times, about the inability to reduce chemistry to an inadequate physics until the 1930s, to take two significant moments from the history of science.

---

[53] Thackray, *op. cit.*, p. 190. Priestley's reasons for welcoming "this extreme development of the Newtonian position" were primarily theological, Thackray concludes.

[54] Yolton, *Thinking Matter*, p. 114.

[55] *Ibid.*, p. 125. See chapters V, VI for discussion. Yolton writes that "There was no British La Mettrie," but that exaggerates La Mettrie's contribution, I believe. See note 51.

A common objection today is that such ideas invoke an unacceptable form of "radical emergence," unlike the emergence of liquids from molecules, where the properties of the liquid can in some reasonable sense be regarded as inhering in the molecules. In Nagel's phrase, "we can *see* how liquidity is the logical result of the molecules 'rolling around on each other' at the microscopic level," though "nothing comparable is to be expected in the case of neurons" and consciousness.[56] Also taking liquidity as a paradigm, Galen Strawson argues extensively that the notion of emergence is intelligible only if we interpret it as "total dependence": if "some part or aspect of *Y* [hails] from somewhere else," then we cannot say that *Y* is "emergent from *X*." We can speak intelligibly about emergence of *Y*-phenomena from non-*Y* phenomena only if the non-*Y* phenomena at the very least are "somehow *intrinsically suited* to constituting" the *X*-phenomena; there must be "something about *X*'s nature in virtue of which" they are "so suited." "It is built into the notion of emergence that emergence cannot be brute in the sense of there being no reason in the nature of things why the emerging thing is as it is." This is Strawson's *No-Radical Emergence Thesis,* from which he draws the panpsychic conclusion that "experiential reality cannot possibly emerge from wholly and utterly non-experiential reality." The basic claim, which he highlights, is that "If it really is true that *Y* is emergent from *X* then it must be the case that *Y* is in some sense wholly dependent on *X* and *X* alone, so that all features of *Y* trace intelligibly back to *X*." Here "intelligible" is a metaphysical rather than an epistemic notion, meaning "intelligible to God": there must be an explanation in the nature of things, though we may not be able to attain it.[57]

Priestley, it seems, would reject Nagel's qualms while accepting Strawson's formulation, but without drawing the panpsychic conclusion. It should be noted that the molecule-liquid example, commonly used, is not a very telling one. We also cannot conceive of a liquid turning into two gases by electrolysis, and there is no intuitive sense in which the properties of water, bases, and acids inhere in Hydrogen or Oxygen or other atoms. Furthermore, the whole matter of conceivability seems to be irrelevant, whether it is brought up in connection with the effects of motion that Newton and Locke found inconceivable, or the irreducible principles of chemistry, or mind-brain relations. There is something about the nature of Hydrogen and Oxygen "in virtue of which they are intrinsically suited to constituting water," so the sciences discovered after long labors, providing reasons "in the nature of things why the emerging thing is as it is." What seemed "brute emergence" was assimilated into science as ordinary emergence—not, to be sure, of the liquidity variety, relying on conceivability. I see no strong reason why matters should necessarily be different in the case of experiential and nonexperiential reality, particularly given our ignorance of the latter, stressed from Newton and Locke to Priestley, developed by Russell, and arising again in recent discussion.

---

[56] Nagel, *op. cit.*

[57] Strawson, "Realistic Monism" and "Reply," in Strawson, *op. cit.* Printers errors corrected (Strawson, p.c.). See essays in this volume for further discussion.

Priestley then considers the claim that mind "cannot be material because it is influenced by reasons." To this he responds that since "reasons, whatever they may be, do ultimately move matter, there is certainly much less difficulty in conceiving that they may do this in consequence of their being the affection of some material substance, than upon the hypothesis of their belonging to a substance that has no common property with matter"—not the way it would be put today, but capturing essentially the point of contemporary discussion leading some to revive panpsychism. But contrary to the contemporary revival,[58] Priestley rejects the conclusion that consciousness "cannot be annexed to the whole brain as a system, while the individual particles of which it consists are separately unconscious." That "A certain quantity of nervous system is necessary to such complex ideas and affections as belong to the human mind; and the idea of self, or the feeling that corresponds to the pronoun I," he argues, "is not essentially different from other complex ideas, that of our country for example." Similarly, it should not perplex us more than the fact that "life should be the property of an entirely animal system, and not the separate parts of it" or that sound cannot "result from the motion of a single particle" of air. We should recognize "that the term self denotes that substance which is the seat of that particular set of sensations and ideas of which those that are then recollected make a part, as distinct from other substances which are the seat of similar sets of sensations and ideas": and "It is high time to abandon these random hypotheses, and to form our conclusions with respect to the faculties of the mind, as well as the properties and powers of matter, by an attentive observation of facts and cautious inferences from them," adopting the Newtonian style of inquiry while dismissing considerations of common-sense plausibility. That seems to be a reasonable stance.

Priestley urges that we also dismiss arguments based on "vulgar phraseology" and "vulgar apprehensions," as in the quest for an entity of the world picked out by the term *me* when I speak of "my body," with its hint of dualism. "According to this merely verbal argument," Priestley observes, "there ought to be something in man besides all the parts of which he consists," something beyond both soul and body, as when "a man says I devote my soul and body," the pronoun allegedly denoting something beyond body and spirit that "makes the devotion." In Rylean terms, phrases of common usage may be "systematically misleading expressions," a lively concern at the time, based on a centuries-old tradition of inquiry into the ways surface grammatical form disguise actual meaning. Like Priestley, Thomas Reid argued that failure to attend "to the distinction between the operations of the mind and the objects of these operations" is a source of philosophical error, as in interpreting the phrase "I have an idea" on the model of "I have a diamond," when we should understand it to mean something like "I am thinking." In an earlier discussion, the Encyclopedist César Chesneau du Marsais, using the same and many other examples, warned against the error of taking nouns to be "names of real objects that exist independently of our thought." The language, then, gives no license for supposing that such words as "idea," "concept," "image" stand for "real objects," let alone "perceptible objects."[59] For similar reasons, Priestley argues that "Nothing surely can be inferred from such phraseology as ["my body"], which, after all, is only derived from vulgar apprehensions."

---

[58] *Ibid.*, Strawson and commentary.

The need to resist arguments from "vulgar apprehensions" holds more broadly: for such phrases as "my thoughts," "my dreams," "my spirit," even "my self," which is different from myself (= me, even though in another sense, I may not be myself these days). When John thinks about himself, he is thinking about John, but not when he is thinking about his self; he can hurt himself but not his self (whatever role these curious entities play in our mental world). There is a difference between saying that his actions are betraying his true (authentic, former) self and that he is betraying himself, and "thine own self" indicates a more essential characteristic than "thyself." Inquiry into manifold questions like these, while entirely legitimate and perhaps enlightening, is concerned with the "operations of the mind," our modes of cognition and thought, and should not be misinterpreted as holding of the "real objects that exist independently of our thought."

---

[59] See my *Aspects of the Theory of Syntax* (Cambridge: MIT, 1965), pp. 199–200, and for much more extensive discussion, *Cartesian Linguistics*. On the accuracy of interpretations of the empiricist theory of ideas by Reid and others, see Yolton, *Perceptual Acquaintance from Descartes to Reid* (Minneapolis: Minnesota UP, 1984), chapter 5.

The latter is the concern of the natural sciences, and I take it also to be the prime concern of the tradition reviewed here.

The operations of the mind doubtless accommodate the thesis that "I am not identical to my body," a core assumption of substance dualism, Stephen Yablo proposes.[60] He suggests further that "substance dualism...has fallen strangely out of view," perhaps "because one no longer recognizes 'minds' as entities in their own right, or 'substances'," though "*selves*—the things we refer to by use of 'I'—are surely substances, and it does little violence to the intention behind mind/body dualism to interpret it as a dualism of bodies and selves." In the tradition I am following here, it is *matter* that has lost its presumed status, and not "strangely." It is also by no means clear, as just noted, that by use of the first person pronoun (as in "I pledge to devote my body and my soul"), or the name "John," we refer to *selves.* But truth or falsity aside, an argument would be needed to show that in using such words we refer (or even take ourselves to be referring) to real constituents of the world that exist independently of our modes of thought. An alternative, which seems to me more plausible, is that these topics belong not to natural science but rather to a branch of ethnoscience, a study of how people think about the world, a very different domain. For natural science, it seems hard to improve on Priestley's conclusion: that Locke's suggestion was fundamentally accurate and that properties "termed mental" reduce to "the organical structure of the brain"—though in ways that are not understood, no great surprise when we consider the history of even the core hard sciences, like chemistry.

As noted above, with the collapse of the traditional notion of body (and so on), there are basically two ways to reconstitute some problem that resembles the traditional mind-body problem: define *physical,* or set the problem up in other terms, such as those that Priestley anticipated.

The first option is developed by Galen Strawson in an important series of publications.[61] Unlike many others, he does give a definition of "physical," so that it is possible to formulate a physical-nonphysical problem. The physical is "any sort of existent [that is] spatio-temporally (or at least temporally) located)." The physical includes "experiential events" (more generally mental events), and permits formulation of the question of how experiential phenomena can be physical phenomena—a "mind-body problem," in a post-Newtonian version. Following Eddington and Russell, and earlier antecedents, notably

---

[60] Yablo, "The Real Distinction between Mind and Body," *Canadian Journal of Philosophy,* Supplementary Volume XVI (1990): 149–201.

[61] Quotes here from "Realistic Monism" and "Reply," in Strawson, *op. cit.*

CENTRAL LIBRARY UNIVERSITY OF CALCUTTA

Priestley, Strawson concludes that "physical stuff has, in itself, 'a nature capable of manifesting itself as mental activity,' that is, as experience or consciousness."

That much seems uncontroversial, given the definitions along with some straightforward facts. But Strawson intends to establish the much stronger thesis of *micropsychism* (which he identifies here with *panpsychism*): "at least some ultimates are intrinsically experience-involving." The crucial premise for that further conclusion, as Strawson makes explicit, is the *No-Radical Emergence Thesis*, already discussed, from which it follows that "experiential reality cannot possibly emerge from wholly and utterly non-experiential reality," a metaphysical issue, not an epistemic one. Strawson interprets Eddington's position to be *micropsychism*, citing his observation that it would be "rather silly to prefer to attach [thought] to something of a so-called 'concrete' nature inconsistent with thought, and then to wonder where the thought comes from," and that we have no knowledge "of the nature of atoms that renders it all incongruous that they should constitute a thinking object." This however appears to fall short of Strawson's micropsychism/panpsychism. Rather, Eddington seems to go no farther than Priestley's conception, writing that nothing in physics leads us to reject the conclusion that an "assemblage of atoms constituting a brain" can be "a thinking (conscious, experiencing) object." He does not, it seems, adopt the No-Radical Emergence Thesis that is required to carry the argument beyond to Strawson's conclusion. Russell too stops short of this critical step, and Priestley explicitly rejects it, regarding radical emergence as normal science. Textual interpretation aside, the issues seem fairly clearly drawn.

The second option is pursued by Daniel Stoljar, who has done some of the most careful work on physicalism and variants of the "mind-body problem." He does offer some answers to the question of what it means to say that something is *physical*[62]—a question which, he notes, has not received a great deal of attention in the literature, though "Without any understanding of what the physical is, we can have no serious understanding of what physicalism is." The answers he offers are not too convincing, I think he would agree, but he argues that it does not matter much: "we have many concepts that we understand without knowing how to analyze," and "the concept of the physical is one of the central concepts of human thought." The latter comment is correct, but only with regard to the common-sense concept of the mechanical philosophy, long ago undermined. The former is correct too, but it is not clear that we want to found a serious philosophical position on a concept that we think we understand intuitively but cannot analyze, particularly when a long history reveals that such common-sense understanding can often not withstand serious inquiry. But Stoljar's more fundamental reason for not being too concerned with characterizing the "physical" is different: the issues, he argues, should be shifted to epistemological terms, not seeking reduction to *the physical*, but taking physicalism to be only the "background metaphysical assumption against which the problems of philosophy of mind are posed and discussed." Thus "when properly understood, the problems that philosophers of mind are interested in are not with the framework [itself], and to that extent are not metaphysical."

---

[62] Quotes in this paragraph from Stoljar, "Physicalism," *Stanford Encyclopedia of Philosophy* (2001).

Stoljar suggests that "the problem mainly at issue in contemporary philosophy is distinct *both* from the mind-body problem as that problem is traditionally understood *and* from the problem as it is, or might be, pursued in the sciences"; a qualification, I think, is that the traditional problem, at least from Descartes through Priestley (taking his work to be the culmination of the post-Newtonian reaction to the traditional problem), can plausibly be construed as a problem within the sciences. The traditional questions "we may lump together under the heading 'metaphysics of mind'," but contemporary philosophy Stoljar takes to be concerned with "epistemic principles" and, crucially, "*the logical problem of experience.*" It might be true that "the notion of the physical fails to meet minimal standards of clarity," he writes, but such matters "play only an illustrative or inessential role in the logical problem," which can be posed "even in the absence of...a reasonably definite conception of the physical."[63]

The logical problem arises from the assumption that (1) there are experiential truths, while it seems plausible to believe both that (2) every such truth is entailed by (or supervenes on) some nonexperiential truth and (3) not every experiential truth is entailed (or supervenes on) some nonexperiential truth. Adopting (1) and (2) (with a qualification, see below), the crucial question is (3). As already discussed, following a tradition tracing back to Newton and Locke, Priestley sees no reason to accept thesis (3): our "very ignorance" of the properties of post-Newtonian *matter* cautions us not to take this step. In Russell's words (which Stoljar cites), experiential truths "are not known to have any intrinsic character which physical events cannot have, since we do not know of any intrinsic character which could be incompatible with the logical properties that physics assigns to physical events." From these perspectives, then, the logical problem does not arise.[64]

---

[63] Stoljar, *Ignorance and Imagination*, pp. 56, 58.

Stoljar's solution to the logical problem, the new "mind-body problem," is similar to the stance of Priestley and Russell, even if put somewhat differently. It is based on his "ignorance hypothesis, according to which we are ignorant of a type of experience-relevant nonexperiential truth," so that the "logical problem of experience" unravels on epistemic grounds.[65] He suggests elsewhere that "the radical view ... that we are ignorant of the nature of the physical or non-experiential has the potential to completely transform philosophy of mind."[66] In Strawson's formulation, the (sensible) line of thought that was well understood up to a half century ago "disappeared almost completely from the philosophical mainstream [as] analytical philosophy acquired hyperdualist intuitions even as it proclaimed its monism. With a few honorable exceptions it out-Descartesed Descartes (or 'Descartes' [that is, the constructed version]) in its certainty that we know enough about the physical to know that the experiential cannot be physical."[67]

The qualification with regard to (2) is that we cannot so easily assume that there are nonexperiential truths; in fact the assumption may be "silly," as Eddington put it. Some physicists have reached such conclusions on quantum-theoretic grounds. The late John Wheeler argued that the "ultimates" may be just "bits of information," responses to queries posed by the investigator. "The actual events of quantum theory are experienced increments in knowledge" (H.P. Stapp).[68] Russell's three grades of certainty suggest other reasons for skepticism. At least, some caution is necessary about the legitimacy even of the formulation of the "logical problem."

Stoljar invokes the ignorance hypothesis in criticizing C.D. Broad's conclusions about irreducibility of chemistry to physics, a close analogue to the Knowledge Argument, he observes. He concludes that Broad was unaware "that chemical facts follow from physical facts," namely the quantum-theoretic facts.[69] But putting the matter that way is somewhat misleading. What happened is that physics radically changed with the quantum-theoretic revolution, and with it the notion of "physical facts." A more appropriate formulation, I think, is to recognize that post-Newton, the concept "physical facts" means nothing more than what the best current scientific theory postulates, hence should be seen as a rhetorical device of clarification, adding no substantive content. The issue of physicalism cannot be so easily dispensed with. Like Marx's old mole, it keeps poking its nose out of the ground.

---

[64] Stoljar, *Ignorance and Imagination*, pp. 17ff., chapter 2, pp. 56f., 104. Stoljar understands the "traditional problem" to be derived from the *Meditations* (45), hence not a problem of the sciences. But though a conventional reading, it is questionable, for reasons already discussed.

[65] *Ibid.*, chapter 4.

[66] Stoljar, in Strawson, *op. cit.*

[67] Strawson, "Realistic Monism," note 21.

[68] Wheeler, *At Home in the Universe* (American Institute of Physics, 1994), Volume 9 of *Masters of Modern Physics*. Stapp, in Strawson, *op. cit.*

[69] Stoljar, *Ignorance and Imagination*, p. 139.

There are also lesser grades of mystery, worth keeping in mind. One of particular interest to humans is the evolution of their cognitive capacities. On this topic, evolutionary biologist Richard Lewontin has argued forcefully that we can learn very little, because evidence is inaccessible, at least in any terms understood by contemporary science.[70] For language, there are two fundamental questions in this regard: first, the evolution of the capacity to construct an infinite range of hierarchically structured expressions interpretable by our cognitive and sensorimotor systems; and second, the evolution of the atomic elements, roughly word-like, that enter into these computations. In both cases, the capacities appear to be specific to humans, perhaps even specific to language, apart from the natural laws they obey, which may have rather far-reaching consequences, recent work suggests. I think something can be said about the first of these questions, the evolution of the generative mechanisms. One conclusion that looks increasingly plausible is that externalization of language by means of the sensorimotor system is an ancillary process, and also the locus of much of the variety and complexity of language. The evolution of atoms of computation, however, seems mired in mystery, whether we think of them as concepts or lexical items of language. In symbolic systems of other animals, symbols appear to be linked directly to mind-independent events. The symbols of human language are sharply different. Even in the simplest cases, there is no word-object relation, where objects are mind-independent entities. There is no reference relation, in the technical sense familiar from Frege and Peirce to contemporary externalists. Rather, it appears that we should adopt something like the approach of the seventeenth and eighteenth century cognitive revolution, and the conclusions of Shaftesbury and Hume that the "peculiar nature belonging to" the linguistic elements used to refer is not something external and mind-independent. Rather, their peculiar nature is a complex of perspectives involving Gestalt properties, cause-and-effect, "sympathy of parts" directed to a "common end," psychic continuity, and other such mental properties. In Hume's phrase, the "identity, which we ascribe" to vegetables, animal bodies, artifacts, or "the mind of man"—the array of individuating properties—is only a "fictitious one," established by our "cognoscitive powers," as they were termed by his seventeenth century predecessors. That is no impediment to interaction, including the special case of communication, given largely shared cognoscitive powers. Rather, the semantic properties of words seem similar in this regard to their phonetic properties. No one is so deluded as to believe that there is a mind-independent object corresponding to the internal syllable [ba], some construction from motion of molecules perhaps, which is selected when I say [ba] and when you hear it. But interaction proceeds nevertheless, always a more-or-less rather than a yes-or-no affair.[71]

---

[70] Lewontin, "The Evolution of Cognition: Questions We Will Never Answer," in Scarborough and Sternberg, *op. cit.*

There is a lot to say about these topics, but I will not pursue them here, merely commenting that in this case too, there may be merit to Strawson's conclusion that "hyperdualist intuitions" should be abandoned along with the "certainty that we know enough about the physical to know that the experiential cannot be physical," and Stoljar's suggestion that "the radical view" might transform philosophy of mind and language, if taken seriously.

Returning finally to the core example of Cartesian science, human language, Gassendi's advice to seek a "chemicallike" understanding of its internal nature has been pursued with some success, but what concerned the Cartesians was something different: the creative use of language, what Humboldt later called "the infinite use of finite means," stressing *use*.[72]

There is interesting work on precepts for language use under particular conditions—notably intent to be informative, as in neo-Gricean pragmatics—but it is not at all clear how far this extends to the normal use of language, and in any event, it does not approach the Cartesian questions of creative use, which remains as much of a mystery now as it did centuries ago, and may turn out to be one of those ultimate secrets that ever will remain in obscurity, impenetrable to human intelligence.

NOAM CHOMSKY

Massachusetts Institute of Technology

---

[71] See *Cartesian Linguistics*, pp. 94ff., and McGilvray's introduction, on Cartesian and neo-Platonist conceptions of the role of "cognoscitive powers." For review and sources on referring, see *New Horizons*; Shaftesbury, Hume, and forerunners, see Mijuskovic, *op. cit.*

[72] On misunderstandings about this matter, see my "A Note on the Creative Aspect of Language Use," *Philosophical Review*, XLI, 3 (July 1982): 423–34.

GÖDEL, NAGEL, MINDS, AND MACHINES*

Some fifty years ago (1957 to be exact), Ernest Nagel and Kurt Gödel became involved in a contentious exchange about the possible inclusion of Gödel's original work on incompleteness in the book, *Gödel's Proof*, then being written by Nagel with James R. Newman.[1] What led to the conflict were some unprecedented demands that Gödel made over the use of his material and his involvement in the contents of the book—demands that resulted in an explosive reaction on Nagel's part. In the end the proposal came to naught. But the story is of interest because of what was basically at issue, namely their provocative related but contrasting views on the possible significance of Gödel's theorems for minds versus machines in the development of mathematics. That is our point of departure for the attempts by Gödel, and later J.R. Lucas and Roger Penrose, to establish definitive consequences of those theorems, attempts which—as we shall see—depend on highly idealized and problematic assumptions about minds, machines, and mathematics. In particular, I shall argue that there is a fundamental equivocation involved in those assumptions that needs to be reexamined. In conclusion, that will lead us to a new way of looking at how the mind may work in deriving mathematics which straddles the mechanist and anti-mechanist viewpoints.

The story of the conflict between Gödel and Nagel has been told in full in the introductory note by Charles Parsons and Wilfried Sieg to the correspondence between them in Volume V of the Gödel *Collected Works*,[2] so I shall confine myself to the high points.

The first popular exposition of Gödel's incompleteness theorems was published by Nagel and Newman in 1956 in an article entitled "Goedel's Proof" for the *Scientific American*.[3] The article was reprinted soon after in the four volume anthology edited by Newman, *The World of Mathematics: A Small Library of the Literature of Mathematics from A' h-mosé*

---

* Revised text of the Ernest Nagel Lecture given at Columbia University on September 27, 2007. I wish to thank Hannes Leitgeb and Carol Rovane for their helpful comments on an earlier version.

[1] Nagel and Newman, *Gödel's Proof* (New York: University Press, 1958); revised edition edited by Douglas R. Hofstadter (New York: University Press, 2001).

[2] Gödel, *Collected Works, Volume V: Correspondence H–Z*, Solomon Feferman et al., eds., (New York: Oxford, 2003), pp. 135ff.

[3] *Scientific American*, CXCIV (June 1956): 71–86.

0022-362X/09/0604/201–19

the Scribe to Albert Einstein, Presented with Commentaries and Notes.[4] That was an instant best-seller, and has since been reprinted many times. Newman had been trained as a mathematician but then became a lawyer and was in government service during World War II. Endlessly fascinated with mathematics, he became a member of the editorial board of the *Scientific American* a few years after the war.

Nagel had long been recognized as one of the leading philosophers of science in the United States, along with Rudolf Carnap, Carl Hempel, and Hans Reichenbach. Like them, he was an immigrant to the U.S.A., but unlike them he had come much earlier, in 1911 at the age of ten. Later, while teaching in the public schools, Nagel received his bachelor's degree at City College of New York in 1923 and his Ph.D. in philosophy at Columbia University in 1930. Except for one year at Rockefeller University, his entire academic career was spent at Columbia, where he became the John Dewey Professor of Philosophy and was eventually appointed to the prestigious rank of University Professor in 1967. In his philosophical work, Nagel combined the viewpoints of logical positivism and pragmatic naturalism. His teacher at City College had been Morris R. Cohen, and with Cohen in 1934 he published *An Introduction to Logic and Scientific Method*,[5] one of the first and most successful textbooks in those subjects.

Soon after the appearance of Nagel and Newman's article on Gödel's theorems in *The World of Mathematics*, they undertook to expand it to a short book to be published by New York University Press. Moreover, they had the idea to add an appendix which would include a translation of Gödel's 1931 paper on undecidable propositions together with the notes for lectures on that work that he had given during his first visit to Princeton in 1934.[6] Early in 1957 their editor at the Press, Allan Angoff, approached Gödel for permission to use that material. Though Gödel said he liked the Nagel-Newman article very much as a nontechnical introduction to his work, he said that he was concerned with some troublesome mistakes[7] that had been made in it and even more so with the interpretation of his results, so he was reluctant about agreeing to the proposal. To encourage him to accept, Nagel paid a lengthy personal visit to him at the Institute in March of 1957. A week later, Gödel wrote Angoff with three conditions. First,

---

[4] In *The World of Mathematics: A Small Library of the Literature of Mathematics from A'h-mosé the Scribe to Albert Einstein, Presented with Commentaries and Notes*, Volume 3 (New York: Simon and Schuster, 1956), pp. 1668–95.

[5] Nagel and Cohen, *An Introduction to Logic and Scientific Method* (New York: Harcourt Brace, 1934).

[6] Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," *Monatshefte für Mathematik und Physik*, XXXVIII (1931): 173–98; reprinted with facing English translation in Gödel, *Collected Works, Volume I: Publications 1929–1936*, Feferman et al., eds. (New York: Oxford, 1986), pp. 144–95.

[7] See Hilary Putnam, "Review of Nagel and Newman (1958)," *Philosophy of Science* XXVII, 2 (1960): 205–07, for a review of Nagel and Newman's *Gödel's Proof* in which

> I would have to write an introduction to the appendix on the one hand in order to mention advances that have been made after the publication of my papers, [and] on the other hand in order to supplement the considerations given in the book, about the philosophical implications of my results. I am not very well pleased with the treatment of these questions that came out in the Scientific American and in the "World of Mathematics". This, for the most part is not the fault of the authors, because almost nothing has been published on this subject, while I have been thinking about it in the past few years.[8]

A second condition, somewhat of a surprise, was that he share in some way in the royalties of the book. But the most contentious condition was the one he put at the beginning of his letter to Angoff: "In view of the fact that giving my consent to this plan implies, in some sense, an approval of the book on my part, I would have to see the manuscript and the proof sheets of the book, including the appendix." This condition was made even more explicit in a follow-up letter:

> Of course I shall have to see the manuscript of the book before I sign the contract, so that I can be sure that I am in agreement with its content, or that passages with which I don't agree can be eliminated, or that I may express my view about the questions concerned in the introduction (*ibid.*, p. 4).

Gödel also wrote Nagel saying that he had made to Angoff "the same suggestions I mentioned to you in our conversation in Princeton." But when Nagel was shown the correspondence with Angoff, he exploded. He thought the proposal to share royalties was "grasping and unreasonable" yet was prepared to accept that. But what really ticked him off, as he wrote Angoff, was Gödel's number one condition:

> I could scarcely believe my eyes when I read his ultimatum that he is not only to see the manuscript of our essay *before* signing the contract with you, but that he is to have the right to eliminate anything in the essay of which he disapproves. In short, he stipulates as a condition of signing the contract the right of censorship.

---

several errors are identified, the most egregious being the misstatement of Gödel's first incompleteness theorem and Rosser's improvement thereof on p. 91.

[8] Gödel, *Collected Works, Volume IV: Correspondence A–G*, Feferman et al., eds. (New York: Oxford, 2003), p. 1.

> This seems to me just insulting, and I decline to be a party to any such agreement with Gödel .... If [his] conditions were granted, Jim and I would be compelled to make any alterations Gödel might dictate, and we would be at the mercy of his tastes and procrastination for a period without foreseeable end .... Gödel is of course a great man, but I decline to be his slave.[9]

Nagel's fears about how things might go if they agreed to this condition were indeed well founded, but he left it in Angoff's hands to communicate his displeasure with it. In the event, Angoff tried to avoid direct conflict, and wrote Gödel in a seemingly accomodating but ambiguous way. It was not until August 1957 that Nagel wrote Gödel himself making clear his refusal to accept it:

> ... I must say, quite frankly, that your ... stipulation was a shocking surprise to me, since you were ostensibly asking for the *right to censor* anything of which you disapproved in our essay. Neither Mr. Newman nor I felt we could concur in such a demand without a complete loss of self-respect. I made all this plain to Mr. Angoff when I wrote him last spring, though it seems he never stated our case to you. I regret now that I did not write you myself, for I believe you would have immediately recognized the justice of our demurrer... (*ibid.*, pp. 152–53).

This was passed over in silence in Gödel's response to Nagel. In the end, the proposal to include anything by him in the book by Nagel and Newman came to naught, and even the specific technical errors which Gödel could have brought to their attention remained uncorrected. Since this was to be the first popular exposition that would reach a wide audience, Gödel had good reason to be concerned, but by putting the conditions in the way that he did he passed up the chance to be a constructive critic.

As an aside, it is questionable whether Gödel indeed recognized the "justice of [their] demurrer." He made a similar request seven years later, when Paul Benacerraf and Hilary Putnam approached him about including his papers on Russell's mathematical logic and Cantor's continuum problem in their forthcoming collection of articles on the philosophy of mathematics. Gödel feared that Benacerraf and Putnam would use their introduction to mount an attack on his platonistic views, and he demanded what amounted to editorial control of it as a condition for inclusion. In that case, the editors refused straight off, but gave Gödel sufficient reassurance about the nature of its contents for him to grant permission to reprint as requested.[10]

---

[9] Gödel, *Collected Works, Volume V: Correspondence H–Z*, pp. 138–39.

[10] Cf. Gödel, *Collected Works, Volume II: Publications 1938–1974*, Feferman et al., eds. (New York: Oxford, 1990), p. 166.

I. GÖDEL'S CONCERNS, PART I: THE FORMULATION OF THE INCOMPLETENESS THEOREMS

Recall that from his first letter to Angoff, one of the conditions that Gödel put for the proposed appendix to the Nagel and Newman book was that he write an introduction to it "in order to mention advances that have been made after the publication of my papers" as well as "to supplement the considerations, given in the book, about the philosophical implications of my results," with which he was "not very well pleased." As to the first of these, Gödel would have wanted to use the opportunity to put on record what he considered the strongest formulation of his incompleteness theorems. This was signaled in a letter that he composed to Nagel early in 1957 but that was apparently never sent. (A number of letters found in Gödel's *Nachlass* were marked *nicht abgeschickt.*) He there writes that

> [c]onsiderable advances have been made ... since 1934. ...it was only by Turing's work that it became completely clear, that my proof is applicable to *every* formal system containing arithmetic. I think the reader has a right to be informed about the present state of affairs.[11]

What he is referring to, of course, is Alan Turing's analysis in 1937 of the concept of effective computation procedure by means of what we now call Turing machines; Gödel had readily embraced Turing's explication after having rejected earlier proposals by Alonzo Church and Jacques Herbrand. But it was not until eight years after the debacle with Nagel and Newman that Gödel spelled out the connection with formal systems and his own work. That was in a postscript he added to the 1965 reprinting of his Princeton lectures in the volume, *The Undecidable,*[12] edited by Martin Davis:

> In consequence of later advances, in particular of the fact that, due to A. M. Turing's work, a precise and unquestionably adequate definition of the general concept of formal system can now be given, the existence of undecidable arithmetical propositions and the non-demonstrability of the consistency of a system in the same system can now be proved rigorously for *every* consistent formal system containing a certain amount of finitary number theory.
>
> Turing's work gives an analysis of the concept of "mechanical procedure" (alias "algorithm" or "computation procedure" or "finite com-

---

[11] Gödel, *Collected Works, Volume V: Correspondence H–Z,* p. 147.

[12] In Martin Davis, ed., *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions* (Hewlett, NY: Raven, 1965); reproduced in Gödel, *Collected Works, Volume I: Publications 1929–1936,* p. 369.

> binatorial procedure"). This concept is shown to be equivalent with that of a "Turing machine". *A formal system can simply be defined to be any mechanical procedure for producing formulas, called provable formulas.* For any formal system in this sense there exists one in the [usual] sense ... that has the same provable formulas (and likewise vice versa) ... [Italics mine]

As we will see, there is much more to this identification than meets the eye; in fact, in my view it is the source of a crucial misdirection in the minds versus machines disputes that we shall take up below. By comparison, Nagel and Newman distinguish the consequences of the incompleteness theorems for axiomatic systems and for "calculating machines" in their concluding reflections as follows:

> [Gödel's theorems] show that there is an endless number of true arithmetical statements which cannot be formally deduced from any specified set of axioms .... It follows, therefore, that an axiomatic approach to number theory ... cannot exhaust the domain of arithmetical truth ....[13]
>
> Gödel's conclusions also have a bearing on the question whether calculating machines can be constructed which would be substitutes for a living mathematical intelligence. Such machines, as currently constructed and planned, operate in obedience to a fixed set of directives built in, and they involve mechanisms which proceed in a step-by-step manner. But in the light of Gödel's incompleteness theorem, there is an endless set of problems in elementary number theory for which such machines are inherently incapable of supplying answers, however complex their built-in mechanisms may be and however rapid their operations (*ibid.*, p. 1695).

Of course, Gödel could charitably read their informal idea of calculating machines simply to be explicated by the notion of Turing machines, but as we saw above, he would have gone beyond that to stress that there is no essential difference between these results.

The first thing that Gödel was surely reacting to on the philosophical rather than the technical side was the statement by Nagel and Newman that

> [the incompleteness theorems] seem to show that the hope of finding an absolute proof of consistency for any deductive system in which the whole of arithmetic is expressible cannot be realized, if such a proof must satisfy the finitistic requirements of Hilbert's original program (*ibid.*, p. 1694).

What must have annoyed Gödel was that this, together with their further reflections on the significance of the incompleteness theorems concerned matters to which he had given a good deal of thought over the years, in some respects with overlapping conclusions, in others contrary ones, but in all cases in much greater depth and with much greater care and precision. The trouble was that almost none of his thought on these questions had been published. With respect to the significance of the incompleteness theorems for Hilbert's finitist consistency program this was something that he had made only one brief cautious comment about at the end of his 1931 paper on undecidable propositions. But it was a matter he kept coming back to all through his life, as we found when we unearthed unpublished lectures and seminar presentations in his *Nachlass*. The first time his further views on this would begin to come out in print would be in 1958, a year after the imbroglio with Nagel and Newman.[14] Since my main concern in the remainder of this article is with the minds versus machines debate as it relates to the incompleteness theorems, I shall leave that matter at that.

---

[13] Nagel and Newman, "Goedel's Proof," *The World of Mathematics*, p. 1694.

II. GÖDEL'S CONCERNS, PART II: SIGNIFICANCE OF THE INCOMPLETENESS THEOREMS FOR THE MINDS VERSUS MACHINES DEBATE

The above quotations from Nagel and Newman's 1956 version of "Goedel's Proof," all come from their final three paragraphs on the "far reaching import" of the incompleteness theorems. (My guess is that Nagel was largely responsible for their formulation, but I have no specific evidence for that.) These, with some rewording but no essential change in content, were to form the entire last chapter, entitled "Concluding Reflections," of the 1958 version of the book *Gödel's Proof*; it is only the 1956 version that Gödel would have seen, however, and to which he would have reacted and so it is from there that I go on to draw the quotations. We now continue with the remaining parts of these paragraphs that concern the potentialities of human thought versus the potentialities of computing machines, already signaled above in their discussion of "calculating machines":

> It may very well be the case that the human brain is itself a "machine" with built-in limitations of its own, and that there are mathematical

---

[14] That was in the article, "Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes," *Dialectica*, XII (1958): 280–87; reprinted with facing English translation in Gödel, *Collected Works, Volume II: Publications 1938–1974*, pp. 240–51. As a sign of his concern with the issues involved, Gödel worked on a revision of that until late in his life, 1972, "On an Extension of Finitary Mathematics Which Has Not Yet Been Used," also in Gödel, *Collected Works, Volume II*, pp. 271–80. For the full story, see my forthcoming piece, "Lieber Herr Bernays! Lieber Herr Gödel! Gödel on Finitism, Constructivity and Hilbert's Program," in *Horizons of Truth* (Gödel centenary conference, Vienna, April 27–29, 2006).

> problems which it is incapable of solving. Even so, the human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines.[15]

And (from the third paragraph)

> The discovery that there are formally indemonstrable arithmetic truths does not mean that there are truths which are forever incapable of becoming known, or that a mystic intuition must replace cogent proof. It does mean that the resources of the human intellect have not been, and cannot be, fully formalized, and that new principles of demonstration forever await invention and discovery .... Nor do the inherent limitations of calculating machines constitute a basis for valid inferences concerning the impossibility of physico-chemical explanations of living matter and human reason. The possibility of such explanations is neither precluded nor affirmed by Gödel's incompleteness theorem. The theorem does indicate that in structure and power the human brain is far more complex and subtle than any nonliving machine yet envisaged (*ibid.*).

And finally,

> Gödel's work is a remarkable example of such complexity and subtlety. It is an occasion not for dejection because of the limitations of formal deduction but for a renewed appreciation of the powers of human reason (*ibid.*).

Here again Gödel would have reacted with a "been there, done that" annoyance, since he had already laid out his thoughts in this direction fifteen years earlier in what is usually referred to as his Gibbs lecture, "Some Basic Theorems on the Foundations of Mathematics and Their Implications."[16] But once more this is something he had never published, though he wrote of his intention to do so soon after delivering the lecture; in fact it never appeared in his lifetime. After Gödel died, the text languished with a number of other important essays and lectures in his *Nachlass* until it was retrieved by our editorial group for publication in Volume III of the Gödel *Collected Works.*

---

[15] Nagel and Newman, "Gödel's Proof," p. 1695.

[16] Gödel's lecture was the twenty-fifth in a distinguished series set up by the American Mathematical Society to honor the nineteenth century American mathematician, Josiah Willard Gibbs, famous for his contributions to both pure and applied mathematics. It was delivered to a meeting of the AMS held at Brown University on December 26, 1951. See Gödel, *Collected Works, Volume III: Unpublished Essays and Lectures,* Feferman et al., eds. (New York: Oxford, 1995), pp. 304–23.

There are essentially two parts to the Gibbs lecture, both drawing conclusions from the incompleteness theorems. The first part concerns the potentialities of mind versus machines for the discovery of mathematical truths, and it is that part that we should compare with Nagel and Newman's reflections. The second part is an argument aimed to "disprove the view that mathematics is only our own creation," and thus to support some version of platonic realism in mathematics. George Boolos wrote a very useful introductory note to both parts of the Gibbs lecture in Volume III of the Gödel *Works* (*ibid.*, pp. 290–304); more recently I have published an extensive critical analysis of the first part, under the title "Are There Absolutely Unsolvable Problems? Gödel's Dichotomy,"[17] and I shall be drawing on that in the following.

What I call Gödel's dichotomy is the following statement that he highlighted in the first part of the Gibbs lecture:

> *Either ... the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems ...* (*op. cit.*, p. 310, italics Gödel's).

By a *diophantine problem* is meant a proposition of elementary number theory (a.k.a. first order arithmetic) of a relatively simple arithmetical form whose truth or falsity is to be determined; its exact description is not important to us. Gödel showed that the consistency of a formal system is equivalent to a diophantine problem, to begin with by expressing it in the form that no number codes a proof of a contradiction.[18] According to Gödel, his dichotomy is a "mathematically established fact" which is a consequence of the incompleteness theorem. All that he says by way of an argument for it is the following, however:

> [I]f the human mind were equivalent to a finite machine, then objective mathematics not only would be incompletable in the sense of not being contained in any well-defined axiomatic system, but moreover there would exist *absolutely* unsolvable problems ..., where the epithet "absolutely" means that they would be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof the mind can conceive (*ibid.*, italics Gödel's).

By a *finite machine* here Gödel means a Turing machine, and by a *well-defined axiomatic system* he means an effectively specified formal system;

---

[17] *Philosophia Mathematica*, Series III, XIV (2006): 134–52.

[18] In modern terms, consistency statements belong to the class Π_1, that is, are of the form $\forall x R(x)$ with R primitive recursive.

as explained above, he takes these to be equivalent in the sense that the set of theorems provable in such a system is the same as the set of theorems that can be effectively enumerated by such a machine. Thus, to say that the human mind is equivalent to a finite machine "even within the realm of pure mathematics" is another way of saying that what the human mind can *in principle* demonstrate in mathematics is the same as the set of theorems of some formal system. By *objective mathematics* Gödel means the totality of true statements of mathematics, which includes the totality of true statements of first-order arithmetic. The assertion that objective mathematics is incompletable is at first sight simply a consequence of the second incompleteness theorem in the form that for any consistent formal system S containing a certain basic system S_0 of (true) arithmetic, the number-theoretical statement Con(S) that expresses the consistency of S is true but not provable in S.

Examined more closely, Gödel's argument is that if the human mind were equivalent to a finite machine, or—what comes to the same thing—an effectively presented formal system S, then there would be *a true statement that could never be humanly proved*, namely Con(S). So that statement would be *absolutely undecidable* by the human mind, and moreover it would be equivalent to a diophantine statement. *Note however, the tacit assumption that the human mind is consistent*; otherwise, it is equivalent to a formal system in a trivial way, namely one that proves all statements. Actually, Gödel apparently accepts a much stronger assumption, namely that we prove *only* true statements; but for his argument, only the weaker assumption is necessary (together of course with the assumption that the basic system of arithmetic S_0 has been humanly accepted). Also, Gödel's sketch to establish his dichotomy should be modified slightly, as follows: either humanly demonstrable mathematics is *contained* in some consistent formal system S, or not. If it is, then Con(S) is an absolutely undecidable (diophantine) statement. If not, then for each consistent formal system S, there is a humanly provable statement that is not provable in S, that is, the "human mind ... infinitely surpasses the powers of any finite machine."

Note well that Gödel's dichotomy is not a strict one as it stands; Gödel himself asserts that "the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives." This could happen, for example, if the human mind infinitely surpasses finite machines with respect to certain diophantine problems, but not with respect to *all* of them. And if we drop the word 'diophantine' from its statement, it might be that the human mind could settle all arithmetical problems, but not all problems of higher mathematics. In fact, some logicians conjecture that Cantor's Continuum Problem is an absolutely undecidable problem of that type.

Be that as it may, how does Gödel's conclusion differ from that of Nagel and Newman? They speak of calculating machines "as currently constructed and planned, [that] operate in obedience to a fixed set of directives built in, and [that] involve mechanisms which proceed in a step-by-step manner," where Gödel speaks more precisely of Turing machines. Still, to be charitable, I think that is a reasonable interpretation of what Nagel and Newman had in mind. Second, they do not make the connection between formal systems and calculating machines, where Gödel sees these as amounting to the same thing.[19] But the essential difference is that they seem to come down in favor of the first, anti-mechanist, disjunct of the dichotomy when they say that "[by] Gödel's incompleteness theorem, there is an endless set of problems in elementary number theory for which such machines are inherently incapable of supplying answers, however complex their built-in mechanisms may be and however rapid their operations" and "the human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines"; furthermore, in the third paragraph they say that "the resources of the human intellect have not been, and cannot be, fully formalized, and ... new principles of demonstration forever await invention and discovery."

There is a lot of evidence outside of the Gibbs lecture that Gödel was also convinced of the anti-mechanist position as expressed in the first disjunct of his dichotomy. That is supplied, for example, in his informal communication of various ideas about minds and machines to Hao Wang, initially in the book, *From Mathematics to Philosophy*,[20] and then at greater length in *A Logical Journey: From Gödel to Philosophy*.[21] So why didn't Gödel state that outright in the Gibbs lecture instead of the more cautious disjunction in the dichotomy? The reason was simply that he did not have an unassailable proof of the falsity of the mechanist position. Indeed, he there says:

> [It is possible that] the human mind (in the realm of pure mathematics) *is* equivalent to a finite machine that, however, is unable to understand completely its own functioning (*op. cit.*, p. 309).

---

[19] Calculating machines are assimilated more closely to axiomatic systems in the concluding reflections of Nagel and Newman, *Gödel's Proof*, p. 100.

[20] Wang, *From Mathematics to Philiosophy* (New York: Routledge and Kegan Paul, 1974), pp. 324–26.

[21] Wang, *A Logical Journey: From Gödel to Philosophy* (Cambridge: MIT, 1996), especially chapter 6.

And in a related footnote, despite his views concerning the "impossibility of physico-chemical explanations of ... human reason" he says that

> [I]t is conceivable ... that brain physiology would advance so far that it would be known with empirical certainty
>
> 1. that the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing;
> 2. that such and such is the precise anatomical structure and physiological functioning of the part of the brain which performs mathematical thinking (*ibid.*).

And in the next footnote, he says:

> [T]he physical working of the thinking mechanism could very well be completely understandable; the insight, however, that this particular mechanism must always lead to correct (or only consistent) results would surpass the powers of human reason (*ibid.*, p. 310).

Some twenty years later, Georg Kreisel made a similar point in terms of formal systems rather than Turing machines:

> [I]t has been clear since Gödel's discovery of the incompleteness of formal systems that we could not have *mathematical* evidence for the adequacy of any formal system; but this does not refute the possibility that some quite specific system ... encompasses all possibilities of (correct) mathematical reasoning ...
>
> *In fact the possibility is to be considered that we have some kind of nonmathematical evidence for the adequacy of such [a system].*[22]

III. CRITIQUING THE MINDS VERSUS MACHINES DEBATE

I shall call the genuine possibility entertained by Gödel and Kreisel, *the mechanist's empirical defense* (or *escape hatch*) against claims to have *proved* that mind exceeds mechanism on the basis of the incompleteness theorems. The first outright such claim was made by J.R. Lucas in his 1961 article, "Minds, Machines and Gödel."[23] Both Benacerraf and Putnam soon objected to his argument on the basis that Lucas was assuming it is known that one's mind is consistent. Lucas, in response, has tried to shift the burden to the mechanist: "The consistency of the machine is established not by the mathematical ability of the mind, but on the word of the mechanist," a burden that the mechanist can refuse to shoulder by simply citing his empirical defense.[24]

---

[22] Kreisel, "Which Number-theoretic Problems Can Be Solved in Recursive Progressions on Π^1_1 Paths through O?" *Journal of Symbolic Logic*, XXXVII (1972): 311–34, see p. 322; italics added.

[23] Lucas, "Minds, Machines, and Gödel," *Philosophy*, XXXVI (1961): 112–37.

Roger Penrose is the other noted defender of the Gödelian basis for anti-mechanism, most notably in his two books, *The Emperor's New Mind*,[25] and *Shadows of the Mind*.[26] Sensitive to the objections to Lucas, he claimed in the latter only to have proved something more modest (and in accord with experience) from the incompleteness theorems: "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth" (*ibid.*, p. 76). But later in that work after a somewhat involved discussion, he came up with a new argument purported to show that the human mathematician cannot even consistently *believe* that his mathematical thought is circumscribed by a mechanical algorithm (*ibid.*, sections 3.16 and 3.23). Extensive critiques have been made of Penrose's original and new arguments in an issue of the journal *PSYCHE*, to which he responded in the same issue.[27] And more recently, Stewart Shapiro[28] and Per Lindström[29] have carefully analyzed and then undermined his "new argument." But Penrose has continued to defend it, as he did in his public lecture for the Gödel Centenary Conference held in Vienna in April 2006.

Historically, there are many examples of mathematical proofs of what cannot be done in mathematics by specific procedures, for example, the squaring of the circle, or the solution by radicals of the quintic, or the solvability of the halting problem. But it is hubris to think that by mathematics alone we can determine what the human mind can or cannot do in general. The claims by Gödel, Lucas, and Penrose to do just that from the incompleteness theorems depend on making highly idealized assumptions both about the nature of mind and the nature of machines. A very useful critical examination of these claims and the underlying assumptions has been made by Shapiro in his article, "Incompleteness, Mechanism, and Optimism,"[30] among which are the following. First of all, how are we to understand the mathematizing capacity of the human mind, since what is at issue is the producibility of an infinite set of propositions? No one mathematician, whose life is finitely limited, can produce such a list, so either what one is talking about is what the individual mathematician *could do in principle*, or we are talking in some sense about the potentialities of the pooled efforts of the community of mathematicians now or ever to exist. But even that must be regarded as a matter of what can be done *in principle*, since it is most likely that the human race will eventually be wiped out either by natural causes or through its own self-destructive tendencies by the time the sun ceases to support life on earth.

---

[24] Lucas, "Minds, Machines, and Gödel: A Retrospect," in P.J.R. Millican and A. Clark, eds., *Machines and Thought: The Legacy of Alan Turing*, Volume 1 (New York: Oxford, 1996), pp. 103–24.

[25] Penrose, *The Emperor's New Mind* (New York: Oxford, 1989).

[26] Penrose, *Shadows of the Mind* (New York: Oxford, 1994).

[27] Penrose, "Beyond the Doubting of a Shadow," *Psyche*, II, 1 (1996): 89–129; also at http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html.

[28] Shapiro, "Mechanism, Truth, and Penrose's New Argument," *Journal of Philosophical Logic*, XXXII (2003): 19–42.

[29] Lindström, "Penrose's New Argument," *Journal of Philosophical Logic*, XXX (2001): 241–50, and "Remarks on Penrose's 'New Argument'," *Journal of Philosophical Logic*, XXXV (2006): 231–37.

[30] Shapiro, "Incompleteness, Mechanism, and Optimism," *Bulletin of Symbolic Logic*, IV (1998): 273–302.

What about the assumption that the human mind is consistent? In practice, mathematicians certainly make errors and thence arrive at false conclusions that in some cases go long undetected. Penrose, among others, has pointed out that when errors are detected, mathematicians seek out their source and correct them,[31] and so he has argued that it is reasonable to ascribe self-correctability and hence consistency to our idealized mathematician. But even if such a one can correct all his errors, can he know with mathematical certitude, as required for Gödel's claim, that he is consistent?

As Shapiro points out, the relation of both of these idealizations to practice is analogous to the competence/performance distinction in linguistics.

There are two further points of idealization to be added to those considered by Shapiro. The first of these is the assumption that the notions and statements of mathematics are fully and faithfully expressible in a formal language, so that what can be humanly proved can be compared with what can be the output of a machine. In this respect it is usually pointed out that the only part of the assumption that needs be made is that the notions and statements of elementary number theory are fully and faithfully represented in the language of first-order arithmetic, and that among those only diophantine statements of the form $\mathrm{Con}(S)$ for S an arbitrary effectively presented formal system need be considered. But even this idealization requires that statements of unlimited size must be accessible to human comprehension.

Finally to be questioned is the identification of the notion of finite machine with that of Turing machine. Turing's widely accepted explication of the informal concept of effective computability puts no restriction on time or space that might be required to carry out computations. But the point of that idealization was to give the strongest *negative* results, to show that certain kinds of problems cannot be decided by a computing machine, no matter how much time and space we allow. And so if we carry the Turing analysis over to the potentiality of mind in its mathematizing capacity, to say that mind infinitely surpasses any finite machine is to say something even stronger. It would be truly impressive if that could be definitively established, but none of the arguments that have been offered are resistant to the mechanist's empirical defense. Moreover, suppose that the mechanist is right, and that in some reasonable sense mind *is* equivalent to a finite machine: Is it appropriate to formulate that in terms of the identification of what is humanly provable with what can be enumerated by a Turing machine? Isn't the mechanist aiming at something stronger in the opposite direction, namely an explanation of the mechanisms that govern the production of human proofs?

---

[31] Cf. Penrose, "Beyond the Doubting of a Shadow," pp. 137ff.

Here is where I think something new has to be said, something that I already drew attention to in my article on Gödel's dichotomy,[32] but that needs to be amplified. Namely, there is an *equivocation* involved that lies in identifying *how* the mathematical mind works with the totality of *what* it can prove. Again, the difference is analogous to what is met in the study of natural language, where we are concerned with the *way* in which linguistically correct utterances are generated and *not* with the potential totality of *all* such utterances. That would seem to suggest that if one is to consider *any* idealized formulation of the mechanist's position at all, it ought to be of the mind as one *constrained* by the axioms and rules of some effectively presented formal system. Since in following those axioms and rules one has *choices* to be made at each step, *at best* that identifies the mathematizing mind with *the program for a nondeterministic Turing machine*, and *not* with the set of its enumerable statements (even though that can equally well be supplied by a deterministic Turing machine).

One could no more disprove this modified version of the idealized mechanist's thesis than the version considered by Gödel and the others, simply by applying the mechanist's empiricist argument. Nevertheless, it is difficult to conceive of any formal system of the sort with which we are familiar, from Peano Arithmetic (PA) up to Zermelo-Fraenkel Set Theory (ZF) and beyond, actually underlying mathematical thought as it is experienced. And the experience of the mathematical practitioner certainly supports the conclusion drawn by Nagel and Newman that "mathematical proof does not coincide with the exploitation of a formalized axiomatic method," even if that cannot be demonstrated unassailably as a consequence of Gödel's incompleteness theorems.

---

[32] Feferman, "Are There Absolutely Unsolvable Problems? Gödel's Dichotomy."

IV. ONE WAY TO STRADDLE THE MECHANIST AND ANTI-MECHANIST POSITIONS

As I see it, the reason for the implausibility of this modified version of the mechanist's thesis lies in the concept of a formal system S that is currently taken for granted in logical work. An essential part of that concept is that the language of S is fixed once and for all. For example, the language of PA is determined (in one version) by taking the basic symbols to be those for equality, zero, successor, addition, and multiplication ($=, 0, ', +, .$), and that of ZF is fixed by taking its basic symbols to be those for equality and membership ($=, \in$). This forces axiom schemata that may be used in such systems, such as induction in arithmetic and separation in set theory, to be infinite bundles of all possible substitution instances by formulas from that language; this makes metamathematical but not mathematical sense. Besides that, the restriction of mathematical discourse to a language fixed in advance, even if only implicitly, is completely foreign to mathematical practice.

In recent years I have undertaken the development of a modified conception of formal system that does justice to the openness of practice and yet gives it an underlying rule-governed logical-axiomatic structure; it thus suggests a way, admittedly rather speculative, of straddling the Gödelian dichotomy. This is in terms of a notion of *open-ended schematic axiomatic system*, that is, one whose schemata are finitely specified by means of propositional and predicate variables (thus putting the 'form' back into 'formal systems') while the language of such a system is considered to be *open-ended*, in the sense that its basic vocabulary may be expanded to any wider conceptual context in which its notions and axioms may be appropriately applied. In other words, on this approach, *implicit in the acceptance of given schemata is the acceptance of any meaningful substitution instances that one may come to meet*, but which those instances are is not determined by restriction to a specific language fixed in advance.[33]

The idea is familiar from standard presentations of propositional and predicate logic, where we have such axioms as

$$P \wedge Q \rightarrow P \quad \text{and} \quad (\forall x)P(x) \rightarrow P(a),$$

---

[33] Cf. Feferman, "Gödel's Program for New Axioms: Why, Where, How and What?" in *Gödel '96*, P. Hájek, ed., *Lecture Notes in Logic*, VI (1996): 3–22, and Feferman, "Open-ended Schematic Axiom Systems (abstract)," *Bulletin of Symbolic Logic*, XII (2006): 145, and Feferman and Thomas Strahm, "The Unfolding of Non-finitist Arithmetic," *Annals of Pure and Applied Logic*, CIV (2000): 75–96.

and rules of inference such as

$$P, P \rightarrow Q \vdash Q \quad \text{and} \quad P \rightarrow Q(x) \vdash P \rightarrow (\forall x) Q(x) \text{ (for } x \text{ not free in } P\text{)}.$$

We do not conceive of logic as applying to a single subject matter fixed once and for all, but rather to any subject in which we take it that we are dealing with well-defined propositions and predicates; logic is there applied by substitution for the proposition and predicate letters in its axioms and rules of inference. Similarly to allow systems like those for arithmetic and set theory to be applicable no matter what subject matter we happen to deal with, we formulate their basic principles in schematic form such as the *Induction Axiom Scheme*

$$P(0) \wedge (\forall x)[P(x) \rightarrow P(x')] \rightarrow (\forall x)P(x)$$

in the case of arithmetic, and the *Separation Scheme*

$$(\forall a)(\exists b)(\forall x)[x \in b \leftrightarrow x \in a \wedge P(x)]$$

in the case of set theory. (The idea for the latter really goes back to Zermelo's conception of the *Aussonderungsaxiom* as applicable to any "definite" predicate.)

But why, it may be (and is) asked, do I insist on the vague idea of an open-ended language for mathematics? Aren't all mathematical concepts defined in the language of set theory? It is indeed the case that the current concepts of working ("pure") mathematicians are with few exceptions expressible in set theory. But there are genuine outliers. For example a natural and to all appearances coherent mathematical notion whose full use is not set-theoretically definable is that of a category; only so-called "small" categories can be directly treated in that way.[34] Other outliers are to be found on the constructive fringe of mathematics in the schools of Brouwerian intuitionism and Bishop's constructivism[35] whose basic notions and principles are not directly accounted for in set theory with its essential use of classical logic. And it may be argued that there are informal mathematical concepts like those of knots, or infinitesimal displacements on a smooth surface, or of random variables, to name just a few, which may be the subject of convincing mathematical reasoning but that are accounted for in set theory only by some substitute notions that share the main expected properties but are not explications in the ordinary sense of the word. Moreover, the idea that set-theoretical concepts and questions like Cantor's continuum problem have determinate mathematical meaning has been challenged on philosophical grounds.[36] Finally, there is a theoretical argument for openness, even if one accepts the language L of set theory as a determinately meaningful one. Namely, by Tarski's theorem, the notion of truth T_L for L is not definable in L; and then the notion of truth for the language obtained by adjoining T_L to L is not definable in *that* language, and so on (even into the transfinite).

---

[34] Cf. Saunders Mac Lane, *Categories for the Working Mathematician* (Berlin: Springer, 1971), and Feferman, "Categorical Foundations and Foundations of Category Theory," in R.E. Butts and J. Hintikka, eds., *Logic, Foundations of Mathematics and Computability Theory*, Volume I (Dordrecht: Reidel, 1977), pp. 149–65, and Feferman, "Enriched Stratified Systems for the Foundations of Category Theory," in G. Sica, ed., *What Is Category Theory?* (Monza, Italy: Polimetrica, 2006), pp. 185–203.

[35] Cf. Michael Beeson, *Foundations of Constructive Mathematics* (Berlin: Springer, 1985).

Another argument that may be made against the restriction of mathematics to a language fixed in advance is historical. Sustained mathematical reasoning had its origins in ancient Greece some 2500 years ago, and mathematical concepts of number and space have undergone considerable evolution since then. Yet even the earliest results have permanent validity, though not necessarily as originally conceived. While Euclidean geometry is no longer considered to be the geometry of actual space, its consequences as an axiomatic development (suitably refined through the work of Pasch, Hilbert, and others), such as the angle sum theorem and Pythagoras's theorem, are as valid now within that context as in Euclid's time. On the other hand, the origins of number theory as represented in Euclid's *Elements*, including the existence of infinitely prime numbers and the fundamental theorem of arithmetic, retain their direct interpretation and validity. But the language of mathematical practice (often mixed with physical concepts) then and through the many following centuries up to the present obviously cannot be identified with the language of set theory. One thing that would instead account for the continuity of mathematical thought throughout its history is the employment of certain underlying formal patterns, such as those indicated above, that could be instantiated by the evolving concepts that have come to fill out mathematics in the process of its development. And there is no reason to believe that this evolutionary process has come to an end; it would be foolish to believe that only that which can be expressed, say, in the language of set theory, will count as mathematics henceforth.

On this picture, in order to straddle the mechanist/anti-mechanist divide at the level considered here, one will have to identify *finitely many basic forms of mathematical reasoning* which work in tandem to fully constrain and distinguish it. These would constitute the mechanist side of the picture, while the openness as to what counts as a mathematical concept would constitute the anti-mechanist side. The evidence for such would have to be empirical, by showing how typical yet substantial portions of the mathematical corpus are accounted for in those terms while giving special attention to challenging cases. I have suggested steps in that direction,[37] but the program is ambitious and I have only made a start; spelling that out is planned for a future publication. In the meantime, the program itself should be treated as highly speculative, yet—I hope—worthy of serious consideration.

---

[36] Feferman, "Does Mathematics Need New Axioms?" *Bulletin of Symbolic Logic*, VI (2000): 401–13.

Considered more broadly and apart from the tendentious terms of the mechanism/anti-mechanism debate, I think the goal should be to give an informative, systematic account at a theoretical level of how the mathematical mind works that squares with experience. Characterizing the logical structure of mathematics—what constitutes a proof—is just one aspect of that, important as that may be. Other aspects—the ones that are crucial in making the difficult choices that are necessary for the mathematician to obtain proofs of difficult theorems—such as the role of heuristics, analogies, metaphors, physical and geometric intuition, visualization, and so on, have also been taken up and are being pursued in a more or less systematic way by mathematicians, philosophers, and cognitive scientists.[38] Without abandoning a basic naturalist stance, I see no usable reductive account of the mathematical experience anywhere on the horizon at the neurophysiological level let alone more basic physico-chemical levels of the sort contemplated by Nagel and Newman and currently sought by Penrose, among others.

SOLOMON FEFERMAN

Stanford University

---

[37] As indicated in "Are There Absolutely Unsolvable Problems?"

[38] Cf., for example, Efraim Fischbein, *Intuition in Science and Mathematics* (Dordrecht: Reidel, 1987); George Lakoff and Rafael E. Núñez, *Where Mathematics Comes From* (New York: Basic, 2000); Paolo Mancosu, Klaus Frovin Jørgensen, Stig Andur Pedersen, eds., *Visualization, Explanation and Reasoning Styles in Mathematics* (Dordrecht: Springer, 2005); and George Pólya, *Mathematics and Plausible Reasoning*, Volumes 1 and 2 (Princeton: University Press, 1968, 2nd edition) among others. In addition, there is a massive amount of anecdotal evidence for the nonmechanical essentially creative nature of mathematical research; for a small sample with further references, cf., for example, Philip J. Davis, and Reuben Hersh, *The Mathematical Experience* (Boston: Birkhäuser, 1981), and David Ruelle, *The Mathematician's Brain* (Princeton: University Press, 2007), and Benjamin H. Yandell, *The Honors Class: Hilbert's Problems and Their Solvers* (Natick, MA: A.K. Peters, 2002).

WHERE OUR NUMBER CONCEPTS COME FROM

Like most cognitive scientists, I take concepts to be mental symbols. Mental symbols are not all concepts, as there are also sensory representations, motor representations, and perceptual representations. From the perspective of cognitive science, a theory of concepts must specify their format, what computations they enter into, what determines their content, and how they differ from other types of mental symbols. In a recent book,[1] I present case studies of the acquisition of several important domains of conceptual representations, arguing that the details of the acquisition process adjudicate among rival theories of concepts within cognitive science. The case studies bear on the existence and nature of innate concepts, and on the existence and nature of discontinuities in development. Human conceptual development involves the construction of representational resources that go beyond those from which they are built in theoretically interesting ways. As Jerry Fodor[2] has forcefully argued, characterizing these discontinuities and explaining how they are possible is a formidable challenge to cognitive science. Meeting this challenge informs our theories of the human conceptual system.

Here I illustrate the lessons I draw from these case studies by touching on one of them: accounting for the origin of concepts of natural number. Explaining the human capacity for representing natural numbers has been a project in philosophy for centuries[3] and in psychology since its emergence as a scientific discipline during the past century.[4] As natural number is the backbone of all of arithmetic, an understanding how representations of natural number arise provides a good start on a theory of the human capacity for mathematics.

Accounting for the origin of any conceptual repertoire requires specifying the innate building blocks from which the representations are built, and specifying the learning mechanisms that accomplish the feat. Two distinct research programs should be, but often are

---

[1] Carey, *The Origins of Concepts* (New York: Oxford, forthcoming).

[2] Fodor, *The Language of Thought* (Cambridge: Harvard, 1975), and "On the Impossibility of Acquiring 'More Powerful' Structures: Fixation of Belief and Concept Acquisition," in M. Piatelli-Palmerini, ed., *Language and Learning* (Cambridge: Harvard, 1980), pp. 142–62.

[3] For example, John Stuart Mill, *A System of Logic* (New York: Harper and Brothers, 1874).

[4] For example, Jean Piaget, *The Child's Conception of Number*, C.G.a.F.M. Hodgson, trans. (New York: Humanities, 1952).

0022-362X/09/0604/220–54

not, distinguished. In one, the logical program, the building blocks are conceived of as logically necessary prerequisites for the capacity in question. In the case of natural number representations these might include the capacity for carrying out recursive computations, the capacity to represent sets, and various logical capacities, such as those captured in second-order predicate calculus. Sometimes arguments within the logical program seek to specify some necessary computational ability that other animals lack[5] that might explain why only humans have the target conceptual ability. A full account of the building blocks for some representational capacity within the logical program must include all of the necessary ones. A second research program, the ontogenetic program, conceives of the building blocks as specific representational systems out of which the target representational capacity is actually built in the course of ontogenesis or historical development. In the case of number representations these would be the innate representations with numerical content (if any).

These two projects are interrelated, but clearly distinct. The first (characterizing the logical prerequisites for natural number) leads to analyses like those that attempt to derive the Peano-Dedekind axioms from Zermelo-Fraenkel set theory or Frege's proof that attempts to derive these axioms from second-order logic and the principle that if two sets can be put in 1–1 correspondence they have the same cardinal value. Such analyses seek to uncover the structure of the concept of natural number, and certainly involve representational capacities drawn upon in mature mathematical thought, but nobody would suppose that in ontogenesis or historical development people construct the concept of natural number by recapitulating such proofs.[6]

The ontogenetic project (characterizing the actual representational systems from which natural number is built) requires empirical studies of infants, nonhuman animals, young children, and historical records, discovering systems of symbols (both mental and public) that are actually created and used in thought, specifying their format and the computations they support. If such empirically attested representational systems do not have the power to represent the target concept (in this case natural number), the project then becomes one of characterizing successive representational systems that are constructed in the process of arriving at the target, and characterizing the learning mechanisms involved in the construction of each.

---

[5] For example, see M.D. Hauser, Noam Chomsky, and W.T. Fitch's proposal that nonhumans lack the capacity for recursion, "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" *Science*, CCXCVIII (2002): 1569–79.

[6] See Feferman, this volume.

The ontogenetic project does not reduce to the logical one. There is no presumption that the concept of natural number can be defined in terms of the earliest representations with numerical content alone. On the other hand, the ontogenetic project depends upon the logical one for a characterization of the target concept at issue (for example, what I mean here by the concept of natural number is characterized by the Peano-Dedekind axioms), as well as for a characterization of the logical resources drawn upon in the construction process.

Many recent papers illustrate the failure to distinguish the two projects. To take just one example at random: A.M. Leslie, Rochel Gelman, and Randy Gallistel[7] argue that natural number is built from an innate system of representations that provides symbols that approximate the cardinal values of sets (analog magnitude representations, see below), plus an innate representation with the content "one," and an innate capacity to represent the successor function. The first of these, analog magnitude number representations, is a well-studied and well-characterized system of representation found throughout the animal kingdom, as well as in human infants, children, and adults. The second, an innate symbol for "one," is a posited innate representation for which these authors offer no evidence. The third is a logical prerequisite for a concept of natural number; clearly we must have the capacity to implement the successor function somehow if we are to represent natural number. There *may* be representations with the content "one" early in development, and as well as systems of representation that implement the successor function, but in the ontogenetic research tradition, one must actually characterize them (format, computational role), and provide evidence for their existence. Thus, one might (and probably would) agree that representations of "one" and the successor function are logical prerequisites for representations of natural number, and in the logical tradition one could explore how representations of number could be built from such primitives (a fairly easy task, obviously). But the ontogenetic program requires providing evidence for such putative developmental primitives.

Here I report on the current state of the art in accounting for the acquisition of concepts of natural number from the ontogenetic research tradition. I argue for four theses: (1) There are three distinct systems of innate representations with numerical content. (2) None contains any symbols for natural numbers. (3) The ontogenetically earliest representational systems that includes symbols for even a subset of the natural numbers is the count list, when deployed in a way that satisfies the "counting principles" described by Gelman and Gallistel.[8] (4) The learning mechanisms that accomplish the construction of the numeral list representation of natural number include, but are not exhausted by, a form of bootstrapping described by W.V.O. Quine,[9] among others, called here "Quinian bootstrapping."

---

[7] Leslie, Gelman, and Gallistel, "The Generative Basis of Natural Number Concepts," *Trends in Cognitive Science*, XII, 6 (2008): 213–18.

NATURAL NUMBER—A CASE STUDY

Leopold Kronecker famously remarked, "The integers were created by God; all else is man-made."[10] Although not what he meant, his remark can be taken to express a view of the *cognitive foundations* of arithmetical thought. If we replace "God" with "evolution," the position would be that evolution provided us with an innate representational system that expresses cardinal values of sets of individuals, an innate representational system that expresses the concepts captured by the Peano-Dedekind axioms. Obviously, evolution yielded *capacity* to construct (by adulthood) representations of natural number—what is actual is possible. Rather, the nativist position is that there is innate computational machinery that operates on representations of sets of individuals and outputs mental symbols with the content of positive integers.

Many modern cognitive scientists, most notably Gelman and Gallistel, have argued for the continuity of integer representations throughout development.[11] Continuing to interpret Kronecker psychologically, he would be saying that all the rest of mathematics, including the rest of the number concepts (rational, negative, 0, real, imaginary, and so on), was culturally constructed by human beings, requiring the construction of representations with more expressive power than those that implement integer representations.

Kronecker's speculation, so interpreted, is wrong. Evolution did not give man the integers. I now turn to the innate representational systems with numerical content we *can* thank evolution for.

Innate system 1: Analog magnitude representations of number. Human adults, human infants, and nonhuman animals deploy a system of analog magnitude representations of number. Number is represented by a physical magnitude that is roughly proportional to the number of individuals in the set whose cardinal value is at stake. Figure 1 depicts

---

[8] Gelman and Gallistel, *The Child's Understanding of Number* (Cambridge: Harvard, 1978).

[9] Quine, *Word and Object* (Cambridge: MIT, 1960), *Ontological Relativity and Other Essays* (New York: Columbia, 1969), and *Roots of Reference* (New York: Columbia, 1974).

[10] Cited in Hermann Weyl, *Philosophy of Mathematics and Natural Science* (Princeton: University Press, 1949), p. 33.

[11] Gellman and Gallistel, *The Child's Understanding of Number*, and "Preverbal and Verbal Counting and Computation," *Cognition*, XLIV (1992): 43–74.

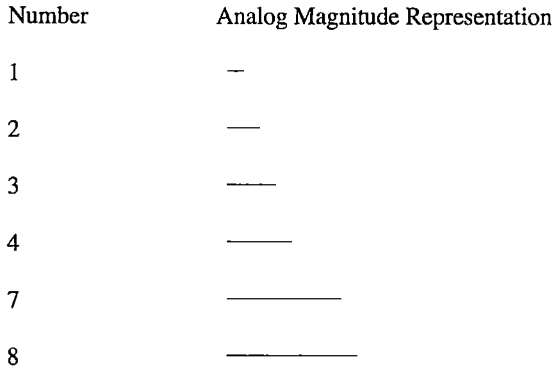


Figure 1: External analog magnitude representation of number in which number is represented by line length.

an external analog magnitude representational system in which lengths represent number. A psychophysical signature of analog magnitude representations is that discriminability accords with Weber's Law. (What I mean by "signature" here is a measurable feature of performance that provides good evidence for some hypothesized underlying representational/computational system). Weber's Law, which holds of representations of many continuous variables (for example, loudness, brightness, length, area, and so on), is that the discriminability of any two magnitudes is a function of their ratio. In the case of number representations, this means that under circumstances where one can discriminate sets of 8 individuals from sets of 16 with a certain reliablility (for example, 75% accuracy), one will be able to discriminate any two sets that stand in a ratio of 1 to 2 with exactly the same accuracy. Weaker consequences of Weber's law can be appreciated by examining the external analogs in Figure 1. It is easy to see that 1 and 2 should be more discriminable than are 7 and 8 (what is called the magnitude effect). Here are 1 and 2: – and —, and here are 7 and 8: ——————— and ————————. Similarly, it should be easier to discriminate 1 from 3 than 2 from 3 (what is called the distance effect). Here are 1 and 3: – and ———, and here are 2 and 3: — and ———. Obviously, both the magnitude and distance effects follow from Weber's law, and indeed, studies of numerical discrimination robustly reveal both magnitude and distance effects.[12]

---

[12] Stanislas Dehaene, *The Number Sense: How the Mind Creates Mathematics* (New York: Oxford, 1997); Gallistel, *The Organization of Learning* (Cambridge: MIT, 1990).

Stanislas Dehaene and Gallistel review the evidence for the long evolutionary history of analog magnitude number representations (*ibid.*). Animals as disparate as pigeons, rats, and nonhuman primates represent number, using the representational system described above. Space precludes a full treatment of this gorgeous literature, so I will present only one series of studies. Elizabeth Brannon and Herb Terrace demonstrated that rhesus macaques are sensitive to number and represent sets differing in cardinal value as numerically ordered. These researchers began with Terrace's previous demonstrations that rhesus macaques can learn to order arbitrary sets of simultaneously presented stimuli. Presented with an array of four objects randomly distributed on a touch screen on each trial, for example, a red circle, a brown table, a black cat, and a blue flower, the monkeys can learn to touch the stimuli in a specified order. Thus, the capacity to represent serial order is itself part of our evolutionary endowment.[13] Brannon and Terrace[14] established that rhesus's number representations are intrinsically ordered. Two monkeys, Rosencrantz and Macduff, were first taught to do the ordered list task, with arbitrary lists such as: first circle, then table, then cat, then flower. Of course, whenever they were shown four new stimuli, they could have no idea what order they were supposed to touch them in, so there was an extended period of trial and error before they learned the order. After Rosencrantz and Macduff became good at that trial and error discovery process, Brannon and Terrace started giving them lists such as those in Figure 2. As you can see, each list consisted of four pictures, containing respectively, sets with 1, 2, 3, and 4 items. In each list, the order the monkeys were supposed to press was 1, 2, 3, 4. Across all the lists, all continuous variables that might have been confounded with number (for example, total surface area) were controlled for. At the beginning, the monkeys treated each list the same as any arbitrary list, requiring extensive trial and error to learn the order called for on that list. But over the course of learning 35 such lists, Rosencrantz and Macduff got faster and faster. This could be because they were becoming ever more efficient at the trial and error strategies for learning whatever arbitrary list the experimenter had in mind, or it could be because they had learned a numerical rule.

---

[13] K.B. Swartz, S. Chen, and Terrace, "Serial Learning by Rhesus Monkeys: Acquisition and Retention of Multiple Four-item Lists," *Journal of Experimental Psychology: Animal Behavior Processes*, XVII, 4 (1991): 396–410; Terrace, L.K. Son, and Brannon, "Serial Expertise of Rhesus Macaques," *Psychological Science*, XIV (2003): 66–73.

[14] Brannon and Terrace, "Ordering of the Numerosities 1 to 9 by Monkeys," *Science*, CCLXXXII (1998): 746–49.

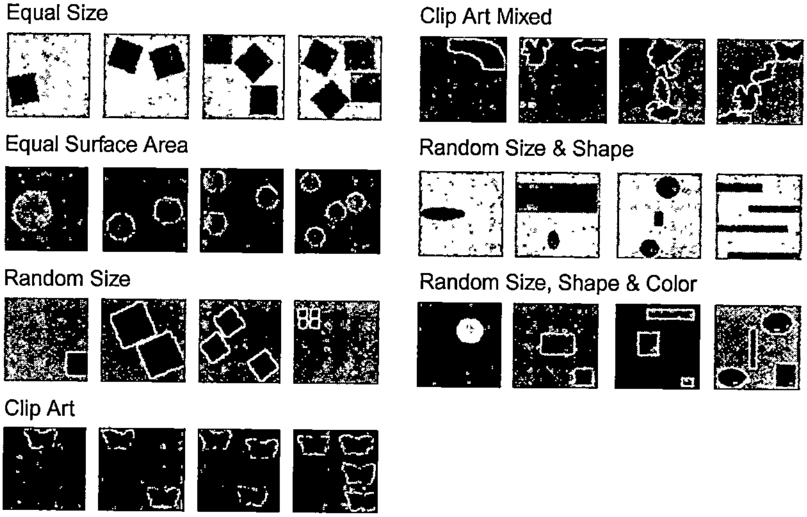


Figure 2: Examples of stimulus lists from Brannon and Terrace.

To decide between these two possibilities, Brannon and Terrace gave Rosencrantz and Macduff 150 trials in which they saw new lists only once, thus preventing any trial and error learning. They did as well as on these lists as on those at the end of the 35 training sets series, where they had seen each list 60 times. They had learned a numerical rule. But which one? "Press 1 object, then 2 objects, then 3 objects, then 4 objects?" Or "press in order of increasing numerical magnitude?" To find out, Brannon and Terrace then presented the monkeys with novel trials involving sets of 4, 6, 7, 8, and 9 items. Now the task was simply to order two stimuli: for example, arrays of 2 versus 4, 3 versus 6, or 5 versus 9. Some included arrays with set sizes within the range of the trained lists, and some were entirely novel set sizes. Each pair was shown only once. Again, both monkeys transferred the rule to the novel pairs of arrays whose cardinal values were outside of the training set. Apparently, they had learned the rule, "touch in order of increasing numerical magnitude."

Analog magnitude representations of number underlie performance on this task. Clear evidence for distance effects were observed—accuracy was a function of the numerical distance between the stimuli (for example, the monkeys' ability to order 6 and 9 was more accurate than their ability to order 7 and 8). Recent data from Brannon's laboratory confirm these generalizations. Monkeys were trained to touch two arrays in numerical order; all training pairs were taken from sets of

1 to 9 elements and over all pairs other variables were controlled as in the Brannon and Terrace studies. After training, sets of 10, 15, 20, and 30 were added. Again, monkeys continued to succeed at the task upon first encountering these larger sets, and performance accorded with Weber's law.[15]

These data support the existence of an evolutionarily ancient representational system in which number is encoded by an analog magnitude proportional to the number of objects in the set. These representations support computations of numerical equivalence and numerical order. There is also evidence that animals can add analog magnitudes.[16] In sum, that these analog magnitude representations are *number* representations is shown by the fact that they track number rather than other dimensions of the sets attended to, and by the fact that numerically relevant computations are defined over them.

In the past eight years four different laboratories have provided unequivocal evidence that preverbal infants form analog magnitude representations of number.[17] The first paper in this flurry of studies is by Fei Xu and Elizabeth Spelke, who solved the problem of how to control for other possible bases of judgment (cumulative surface area, element size, density) in a large number habituation paradigm. The idea underlying habituation studies is simple: one shows a baby stimuli with a common property, one at a time, measuring how long the baby maintains interest. After the baby has become bored, one presents either another stimulus with that property or a stimulus with a contrasting property. If the infant recovers interest (dishabituates) only in the latter case, then we can conclude that the infant was sensitive to the property in question. Xu and Spelke habituated 6-month-old infants to displays containing 8 dots or to displays containing 16 dots

---

[15] J. Cantlon and Brannon, "The Effect of Hetereogeneity on Numerical Ordering in Rhesus Monkeys," *Infancy*, IX, 2 (2006): 173–89.

[16] For example, J.I. Flombaum, J.A. Junge, and Hauser, "Rhesus Monkeys (Macaca Mulatta) Spontaneously Compute Addition Operations over Large Numbers," *Cognition*, XCVII, 3 (2005): 315–25.

[17] Brannon, "The Development of Ordinal Numerical Knowledge in Infancy," *Cognition*, LXXXIII (2002): 223–40; Brannon, S. Abbot, and D.J. Lutz, "Number Bias for the Discrimination of Large Visual Sets in Infancy," *Cognition*, XCIII (2004): B59–68; J.S. Lipton and Spelke, "Origins of Number Sense: Large Number Discrimination in Human Infants," *Psychological Science*, XV, 5 (2003): 396–401, and "Discrimination of Large and Small Numerosities by Human Infants," *Infancy*, V, 3 (2004): 271–90; McCrink and Wynn, "Large-number Addition and Subtraction by 9-Month Old Infants," *Psychological Science*, XV, 11 (2004): 776–81; J. Wood and Spelke, "Infants' Enumeration of Actions: Numerical Discrimination and Its Signature Limits," *Developmental Science*, VIII, 2 (2005): 173–81; Xu and Spelke, "Large Number Discrimination in 6-Month Old Infants," *Cognition*, LXXIV (2000): B1–11; Xu, Spelke, and Goddard, "Number Sense in Human Infants," *Developmental Science*, VIII, 1 (2005): 88–101.

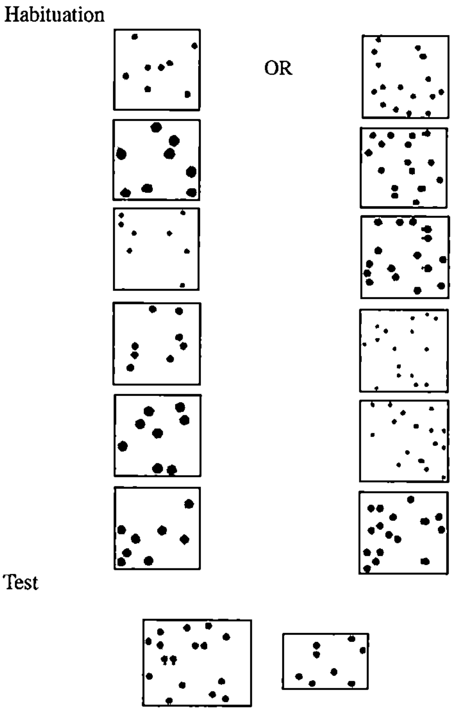


Figure 3: Sample habituation sequences and test stimuli from Xu and Spelke.

(Figure 3). Possible confounds between number and other variables were controlled either by equating the two series of stimuli on those variables, or by making the test displays equidistant from the habituation displays on them. Habituated to 8-dot displays, 7-month-old infants recovered interest when shown the novel 16-dot displays, while generalizing habituation to the novel 8-dot displays. Those habituated to 16-dot displays showed the reverse pattern. Subsequent studies duplicated this design (and the positive result) with 16-dot versus 32-dot comparisons and with 4-dot versus 8-dot comparisons. Thus, the infants showed a sensitivity to cardinal values of sets outside the range of object-tracking mechanisms, when properties of the arrays other than the number of dots each was composed of were strictly controlled for.

That analog magnitude representations support these discriminations is shown by the fact that success is a function of the ratio of the set sizes. In all of the above studies, in which infants succeeded with a 2:1 ratio, they failed in comparisons that involved a 3:2 ratio (that is, they failed to discriminate 8-dot from 12-dot arrays, 16-dot from 24-dot arrays, and 4-dot from 6-dot arrays). Also, these researchers have found that sensitivity improves by 9 months of age. Infants of this age succeed at 3:2 comparisons across a wide variety of absolute set sizes, but fail at 4:3 comparisons.

These baby experiments involve dot arrays. Animals create analog magnitude representations of number of sounds, number of bar presses, and number of key pecks, as well as of sets of visually presented individuals.[18] Infants also represent the cardinal values of sets of individuals that are not visually specified. Jennifer Lipton and Spelke showed that 6-month-old infants discriminate 8 from 16 tones, and also 4 from 8 tones, when continuous variables are controlled in a manner analogous to the Xu and Spelke studies, and they fail to discriminate 8-tone from 12-tone sequences, or 4-tone from 6-tone sequences. Not only do 6-month infants create analog magnitude representations of number in a sequence of tones, their sensitivity to numerical differences is in the same ratio as for arrays of dots (between 2:1, where they succeed, and 3:2, where they fail). Also paralleling the results with dot arrays, by 9-months infants succeed at this latter ratio, distinguishing 8-tone sequences from 12-tone sequences and distinguishing 4-tone sequences from 6-tone sequences. Succeeding at 3:2 ratios, irrespective of set size, infants this age fail at 5:4 ratios, failing to distinguish 5-tone sequences from 4-tone sequences and also 10-tone from 8-tone sequences. Thus at each age sensitivity is a function of the ratios of the number of elements in the sets to be compared, with older infants showing greater sensitivity (3:2 at 9-months; 2:1 at 6-months[19]). Parallel findings are observed in studies of infants' enumeration of events such as jumps of a puppet.[20]

In all of the above studies we can be confident it is number infants are responding to, because every other variable is equated either across the habituation stimuli or across the test stimuli. In these studies, the child's attention is drawn when there is a *different number* of dots, jumps, or tones in a test set from the number in each of the habituation sets, and discrimination follows Weber's law. Thus, the child is using the analog magnitude system to compute numerical equivalence. Of course, if the analog magnitude representations underlying performance in these habituation studies are truly numerical representations, number relevant computations other than establishing numerical equivalence should be defined over them, and indeed this is so.

---

[18] Dehaene, *The Number Sense*; Gallistel, *The Organization of Learning*.

[19] Lipton and Spelke, "Origins of Number Sense," and "Discrimination of Large and Small Numerosities by Human Infants."

[20] Wood and Spelke, "Infants' Enumeration of Actions."

Brannon showed that 11-month-old infants represent numerical order using analog magnitude representations of sets. Infants were habituated to 3 array sequences, always increasing in number by a ratio of 2: 1 (for example, 2, 4, 8; 4, 8, 16; 1, 2, 4). Continuous variables were controlled as in the Xu and Spelke studies. After habituation, infants were shown a novel increasing sequence (3, 6, 12) or a decreasing sequence (12, 6, 3). They remained bored by the the former and dishabituated to the latter. Another group of infants were habituated to decreasing sequences; they generalized habituation to the test sequence that decreased in numerical value and dishabituated to the one that increased.[21]

Finally, Koleen McCrink and Karen Wynn showed that 9-month-olds can manipulate sets of objects in the analog magnitude range to support addition and subtraction.[22] Shown 5 objects moving behind a screen, followed by another 5, infants look longer if the screen is removed revealing a set of 5 than if a set of 10 is revealed. Conversely, if the first set was 10, and 5 objects were seen to leave, infants looked longer if upon the screen's removal 10 were revealed. These objects were each constantly expanding and contracting so it was possible to control for summed continuous variables as a basis of response. In sum, analog magnitude representations of number are available at least by 7 months of age. Preverbal infants represent the approximate cardinal value of sets, and compute numerical equivalence, numerical order, addition, subtraction, and ratios[23] over these representations. Given the ancient evolutionary history of analog magnitude number representations, it is very likely they are the output of innate perceptual analyzers.

In sum, the evidence points to a system of representation in which number is encoded in the brain by some analog symbol that is a linear or logarithmic function of number. It is easy to imagine a system of representation with these properties. For example, R.M. Church and W.H. Meck[24] proposed an accumulator model. Imagine a pulse generator that constantly creates some source of energy, and a gate that can allow energy to flow into a device that registers how much has passed through. If the gate were programmed to open for a fixed amount of time (for example, 250 msec) for each individual in a set, the energy in the accumulator would be a linear function of the number of individuals in the set. Adding could be achieved by combining the energy in two accumulators. Many other models have been proposed, some operating on the individuals in the set in parallel.[25] Considering these models, we can say more about what is represented *explicitly* and what *implicitly* by this system. The symbols themselves are explicit. They are the output of analyzers that take sets of individuals as input and they are available to central processors for a wide variety of computations. But much of the numerical content of this system of representation is implicit. There is no explicit representation of the axioms of arithmetic, no computations of 1–1 correspondence that establish numerical equivalence. Aspects of these principles are implicit in the operation of the input analyzers and in the computations defined over analog magnitudes, but they need not be available for the child to base any decisions on. This case illustrates that once one has a well-confirmed model of some representational system, one can examine that model to establish exactly what is represented and how. Analog magnitudes are explicit symbols for cardinal values of sets, although they only approximate cardinal values. Other numerical content is implicitly embodied in operations that compute over these symbols; that latter knowledge is not symbolized and thus is not input to further computations.

---

[21] Brannon, "The Development of Ordinal Numerical Knowledge in Infancy."

[22] McCrink and Wynn, "Large-number Addition and Subtraction by 9-Month Old Infants."

[23] McCrink and Wynn, "Ratio Abstraction by 6-Month Old Infants," paper presented at the International Conference on Infant Studies, Chicago (2004).

[24] Church and Meck, "The Numerical Attribute of Stimuli," in H.L. Roitblatt, T.G. Bever, and Terrace, eds., *Animal Cognition* (Hillsdale, NJ: Erlbaum, 1984), pp. 445–64.

A second innate system with numerical content: parallel individuation of small sets. Science moves rapidly, and the infant studies reviewed above came relatively late in the history of studies designed to show that infants are sensitive to number. The first studies, some 20 years earlier than Xu's and Spelke's studies on analog magnitude representations, concerned *small* sets—discriminations among sets of 1, 2, and 3 objects. These include many 2 versus 3 habituation studies and Wynn's 1 + 1 = 2 or 1 violation of expectancy studies.[26] In a violation of expectancy study, infants watch an event unfold that either does or does not involve a magic trick. If infants' attention is drawn to the events with impossible outcomes, we conclude that infants' representations of these events involve models that were constrained by the principle violated. In Wynn's famous study, she introduced an object onto a stage, covered it by a screen, and then introduced a second object behind the screen. The screen was then removed to reveal either 2 objects (expected outcome) or 1 or 3 objects (unexpected outcomes). Infants' attention was drawn more to outcomes of 1 or 3 objects than to outcomes of 2, suggesting that infants' representations of these events respected the generalization that 1 + 1 is precisely 2.

---

[25] For example, Dehaene and J.P. Changeux, "Development of Elementary Numerical Abilities: A Neuronal Model," *Journal of Cognitive Neuroscience*, v (1993): 390–407.

[26] S. Antell and D.P. Keating, "Perception of Numerical Invariance in Neonates," *Child Development*, LIV (1983): 695–701; P. Starkey and R. Cooper, "Perception of Numbers by Human Infants," *Science*, CCX (1980): 1033–35; Wynn, "Addition and Subtraction by Human Infants," *Nature*, XXXLVIII (1992): 749–50.

Although some have suggested that analog magnitude number representations underlie success in these experiments,[27] the evidence conclusively implicates a very different representational system.[28] In this alternative representational system, number is only implicitly encoded; there are no symbols for number at all, not even analog magnitude ones. Instead, the representations include a symbol for each individual in an attended set. Thus, a set containing one apple might be represented: "O" (an iconic object file) or "apple" (a symbol for an individual of the kind apple) and a set containing two apples might be represented "O O" or "apple apple," and so forth. These representations consist of one symbol (file) for each individual, and when the content of a symbol is a spatiotemporally determined object, it is called an object file.[29] Infants also create working memory models of small sets of other types of individuals, such as sound bursts or events, and so I shall call the system of representation "parallel individuation" and the explicit symbols within it "individual files." When those individual files are object files, I refer to them as such.

There are many reasons to favor individual file representations over analog magnitude representations as underlying performance in most of the infant small number studies.[30] First, and most important, success on many spontaneous number representation tasks involving small sets do not show the Weber-fraction signature of analog magnitude representations; rather they show the set-size signature of individual file representations. That is, individuals in small sets (sets of 1, 2, or 3) can be represented, and sets outside of that limit cannot, even when the sets to be contrasted have the same Weber-fraction as those small sets where the infant succeeds. Suppose at a given age the set-size ratio an infant is sensitive to is 1:2 (as 6-month-olds are in the large number habituation studies described above). If analog magnitude representations of small sets underlie infants performance, they should succeed at discriminating 1 from 2, 2 from 4, 3 from 6, but fail at discriminating 2 from 3, 3 from 4, or 4 from 6. However, if parallel individuation underlies success, and their limit is three individuals represented in parallel, then they should be able to encode sets of 1, 2, or 3 but should be unable to encode sets of 4 or more.

---

[27] For example, Dehaene, *The Number Sense.*

[28] L. Feigenson and Carey, "Tracking Individuals via Object Files: Evidence from Infants' Manual Search," *Developmental Science*, VI, 5 (2003) 568–84; Feigenson, Carey, and Hauser, "The Representations Underlying Infants' Choice of More: Object Files versus Analog Magnitudes," *Psychological Science*, XIII (2002): 150–56; B.J. Scholl and A.M. Leslie, "Explaining the Infant's Object Concept: Beyond the Perception/Cognition Dichotomy," in Ernest Lepore and Zenon Pylyshyn, eds., *What Is Cognitive Science?* (New York: Oxford, 1999), pp. 26–73; T.J. Simon, "Reconceptualizing the Origins of Number Knowledge: A "Non-numerical" Account," *Cognitive Development*, XII (1997): 349–72; C. Uller, Carey, G. Huntley-Fenner, and L. Klatt, "What Representations Might Underlie Infant Numerical Knowledge," *Cognitive Development*, XIV (1999): 1–36.

[29] D. Kahnemann, A. Triesman, and B. Gibbs, "The Reviewing of Object Files: Object-specific Integration of Information," *Cognitive Psychology*, XXIV (1992): 175–219.

[30] See Carey, *The Origins of Concepts*, for a more thorough review.

The set-size signature of object-file representations is motivated by evidence that even for adults there are sharp limits on the number of object-files that can be simultaneously attended to and held in working memory.[31] This is what generates the prediction that if object-file representations underlie infants' performance in some tasks meant to reflect number representations, then infants should succeed only when the sets being encoded consist of small numbers of objects. Success at discriminating 1 versus 2, and 2 versus 3, in the face of failure with 3 versus 4 or 4 versus 5 is not enough to confirm that object-file representations underlie success, for Weber-fraction differences could equally well explain such a pattern of performance. That is, ratios of 3:4 or 4:5 might exceed the sensitivity of the analog magnitude system at that age. Rather, what is needed is success at 1 versus 2 and perhaps 2 versus 3 in the face of failure at 3 versus 6—failure at the higher numbers when the Weber fraction is the same or even more favorable than that within the range of small numbers at which success has been obtained. This is the set-size signature of individual file representations.

This set-size signature of object-file representations is precisely what is found in some infant habituation studies—success at discriminating 2 versus 3 objects in the face of failure at discriminating 4 versus 6 objects.[32] Similarly, two other paradigms provide vivid illustrations of the set-size signature of object-file representations. In one of them an infant watches as each of two opaque containers, previously shown to be empty, is baited with a different number of graham crackers. For example, the experimenter might put two graham crackers in one container and three in the other. After placement, the parent allows the infant to crawl toward the containers. The dependent measure is which container the baby chooses. The data reflect the set-size signature of parallel individuation. Apparently, three is the limit on parallel individuation in babies. Ten- to 12-month-old infants succeed at 1 versus 2, 2 versus 3, and 1 versus 3; and fail at 3 versus 4, 2 versus 4, and even 1 versus 4.[33] One:four is a more favorable ratio than 2:3, but infants fail at 1 versus 4 comparisons and succeed at 2 versus 3. This pattern of performance shows that analog magnitude representations do not underlie performance on this task; performance does not accord with Weber's law. Note also that 5 crackers are involved in each choice, so the total length of time of placements is equated over these two comparisons.

---

[31] Pylyshyn and R.W. Storm, "Tracking Multiple Independent Targets: Evidence for a Parallel Tracking Mechanism," *Spatial Vision*, III (1998): 179–97; S.J. Luck and E.K. Vogel, "The Capacity of Visual Working Memory for Features and Conjunctions," *Nature*, CCCXC (1997): 279–81.

[32] Starkey and Cooper, "Perception of Numbers by Human Infants."

This is a striking result. Infants could succeed at 1 versus 4 comparisons on many different bases: putting 4 crackers into a bucket takes much longer, draws more attention to that bucket, and so on, yet infants are at chance. Although infants could solve this problem in many different ways, apparently they are attending to each cracker, creating a model of what is in the container that contains one object-file for each cracker. As soon as one of the sets exceeds the limits on parallel individuation, performance falls apart. This finding provides very strong evidence that parallel individuation underlies success on this task.

Convergent data from a second paradigm involving small sets of objects also demonstrate the set-size signature of parallel individuation. The task requires infants to search in a box into which they can reach but not see. When 12- to 14-month-old infants have seen 1, 2, or 3 objects placed into a box, they search for exactly 1, 2, or 3, respectively. But when they have seen 4 objects placed in the box, they are satisfied when they have retrieved only 2 or even only 1. That is, as in the cracker choice experiments, infants distinguish 2 from 3 (see 3 hidden, retrieve 2, expect another in there), but fail to distinguish 4 from 1 (see 4 hidden, retrieve 1, do not search further for any more in there).[34] Performance falls apart when the set to be represented exceeds the limit on parallel individuation of objects, not when the ratio of objects exceeds some limit. Again, 4:1 is a more favorable ratio than 3:2, yet infants search for additional objects having seen 3 placed into the box and having retrieved only 2, but fail to search for additional objects having seen 4 placed into the box and having retrieved only 1.

That infants' performance shows the set-size signature of parallel individuation rather than the constant Weber ratio signature of

---

[33] Feigenson and Carey, "On the Limits of Infants' Quantification of Small Object Arrays," *Cognition*, XCVII, 3 (2005): 295–313; Feigenson, Carey, and M. Hauser, "The Representations underlying Infants' Choice of More: Object Files vs. Analog Magnitudes," *Psychological Science*, XIII, 2 (2002): 150–56.

[34] Feigenson and Carey, "Tracking Individuals via Object Files," and "On the Limits of Infants' Quantification of Small Object Arrays."

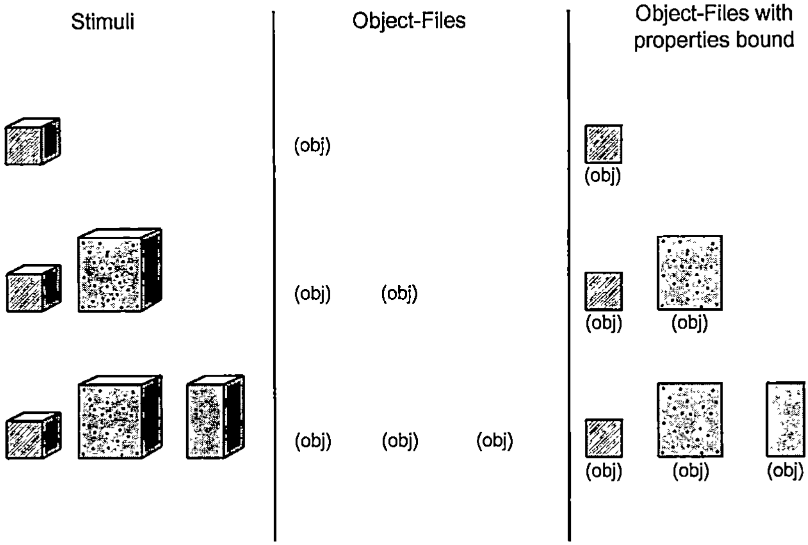


Figure 4: Two versions of the memory structures that might subserve parallel individuation of small sets of objects. In one, each object is represented by an object-file that abstracts away from specific features (OBJ). In the other, each object is represented by an object-file on which shape, color, texture and spatial extent features have been bound.

analog magnitudes shows that analog magnitude representations cannot subserve performance on the small number tasks described above. Rather, these arrays are being represented in short-term memory models that consist of one symbol for each individual in the small set being represented. Although some describe what I am calling the system of parallel individuation a "small number system," that is a misleading name. The purpose of parallel individuation is to create working memory models of small sets of individuals, in order to represent spatial, causal, and other relations among them. Unlike analog magnitude number representations, the parallel individuation system is not a dedicated number representation system. Far from it. The symbols in the parallel individuation system explicitly represent individuals. Figure 4 depicts several different individual file representations of two boxes. In none of these alternatives is there a symbol that has the content "two"; rather, the symbols represent the boxes. The whole model {box box} represents two boxes, of course, but the content "two" is only implicit. Furthermore, the quantitative calculations over parallel individuation models in working memory often privilege continuous variables (such as total event energy, total contour length, total surface area) over numerical equivalence. So why I am discussing these models in a paper on the origin of natural number?

Parallel individuation models are shot through with numerical content, even though that numerical content is merely implicit in the computations that pick out and index small sets to represent, that govern the opening of new individual files, that update working memory models of sets as individuals are added or subtracted, and that compare sets on numerical criteria. The creation of a new individual file requires principles of individuation and numerical identity; models must keep track of whether this object or jump, seen now, is the same *one* as that object seen before, or this sound just heard, is the same *one* as that just heard previously. The decision the system makes dictates whether an additional individual file is established, and this guarantees that a model of a set of three boxes will contain three box symbols. Computations of numerical identity are (as their name says) numerical computations. Also, the opening of a new individual file in the presence of other active files provides an implicit representation of the process of adding one to an array of individuals. Finally, and this is a very important point for what is to come, working memory models of *two* sets of individuals can be simultaneously maintained, and when individual-file models are compared on the basis of 1–1 correspondence, the computations over these symbols establish numerical equivalence and numerical order. Remember the cracker choice experiments: infants succeed at 3 versus 2 comparisons, even though the total number of individuals involved in these events is 5. They fail when the number of individuals in either set exceeds the limits of parallel individuation.

Several recent studies show that infants carry out computations on these models that determine numerical equivalence and numerical order. To this end, infants compute 1–1 correspondence between representations of small sets held in short-term memory. I will use one of these studies for illustration. Lisa Feigenson and I studied the nature of the computations that determined whether infants searched for additional objects in the box in the studies described above. We carried out a version of this study in which infants saw 2 small objects, for example, cars, placed into the box, one at a time. We then gave them the box, and they reached in and retrieved a car—either one of the cars they had seen, or a car that was twice the surface area and 4 times the volume but was otherwise identical to the cars they saw hidden. Infants showed by their subsequent search that they expected exactly one more in either case. They were oblivious to the cumulative continuous variables; their reaches were guided by how many objects they represented in the box.[35] They must have been computing 1–1 correspondence between the individuals in their model of what was in the box and their representations of what they had so far retrieved from it.

I am often asked why infants do not draw upon analog magnitude representations in these tasks. Analog magnitudes are defined for small set sizes, certainly for adults[36] and probably for infants.[37] I do not know for sure, but aspects of the stimulus presentation that draw attention to objects as individuals (a small set, with individuals moving independently of each other and/or with each individual object very different from the others in the set) apparently leads infants to attempt to model the situation with parallel individuation. Presenting large sets, outside of the range of parallel individuation, often with little time to attend successively to each separate individual, encourages analog magnitude representations. As cognitive scientists, we do not attempt to explain why any particular thought is entertained at any given moment; rather we attempt to isolate distinct systems of representation, so as to characterize them. The studies I have described implicate two separable systems of representation with numerical content, each with very different properties.

A third innate system of representation with numerical content: natural language quantifiers. Please dwell upon the experiments that reveal the set-size signature of parallel individuation. Infants choose a set of 3 crackers over a set of 2 crackers or a single cracker; and shown 3 balls placed into a box, having retrieved 2 or 1 of these, they search for the remaining ones. However, they are at chance in a comparison of 1 and 4 crackers, and shown 4 balls placed into a box, they are satisfied after retrieving only 1. These are very counterintuitive phenomena. Not only are infants failing to encode the set of four as approximately four (using analog magnitudes), they are failing to encode it as "plural," for if they had done so, they would represent it as more than one.

The infants in these studies are 12 to 14 months of age; the failures at 1 versus 4 comparisons in the box-search task are also observed at 16, 18, and 20 months of age.[38] In spite of these failures, we now know that there are circumstances in which prelinguistic infants and nonhuman primates reveal representations of the singular-plural distinction. In the above studies, the individuals move independently of each other, encouraging the infants to deploy parallel individuation. If sets of objects move as coherent wholes (for example, glued to a platform, such that the items move together), young infants and rhesus macaques distinguish singletons from sets of more than one and fail to distinguish among plural sets of different numerosities, at least for small sets (for example, 2, 3, 4, and 5.)[39] Again, we do not know why monkeys and infants do not draw on analog magnitudes on these tasks, but the data indicated that they do not, whereas their pattern of behavior reflects a categorical distinction between singletons, on the one hand, and pluralities, on the other.

---

[35] Feigenson and Carey, "Tracking Individuals via Object Files."

[36] Cordes, Gelman, and Gallistel, "Variability Signatures Distinguish Verbal from Nonverbal Counting for Both Large and Small Numbers," *Psychological Bulletin and Review*, VIII (2001): 698–707.

[37] Brannon, "The Development of Ordinal Numerical Knowledge in Infancy."

[38] D. Barner, D. Thalwitz, Wood, and Carey, "On the Relation between the Acquisition of Singular-plural Morpho-syntax and the Conceptual Distinction between One and More than One," *Developmental Science*, x (2007): 365–73.

Infants learning English come to comprehend some explicit markers of the English singular/plural distinction (for example, "are some" versus "is a") by 22 months of age. The available evidence suggests that the singular determiner "a" is given the adult analysis as soon as it is acquired, referring to a single individual of a kind (rather than to a proper noun). By this age infants distinguish "this is a blicket" from "this is Blicket," as well as "there is a doll" from "there are some dolls." The ease of mastering determiners and other quantifiers, as well as evidence for prelinguistic availability of the singularity/plurality distinction, suggests that the meanings that underlie the quantifiers of natural language are also part of the human innate endowment.

What are those meanings? Semanticists such as G. Link,[40] Gennaro Chierchia,[41] and others show that the meanings of quantifiers are formulated over representations of atoms (individuals), and the semilattice of sets that can be constructed from a domain of atoms. Unlike the two representational systems described above, little is known about exactly how (format) the meanings of quantifiers are represented in the mind.

Conclusions: The building blocks for representations of natural number. The three systems of representation described above are the only innate systems of representation with numerical content for which we currently have any evidence. This does not mean, of course, that there might not be others, not yet discovered or described. Perhaps, but I doubt it.

---

[39] Barner, Wood, Hauser, and Carey, "Evidence for a Non-linguistic Distinction between Singular and Plural Sets in Rhesus Monkeys," *Cognition*, CVII, 2 (2008): 603–22; Barner, Thalwitz, Wood, and Carey, unpublished data.

[40] Link, "Generalized Quantifiers and Plurals," in P. Gärdenfors, ed., *Generalized Quantifiers* (Boston: Reidel, 1987), pp. 151–80.

[41] Chierchia, "Plurality of Mass Nouns and the Notion of 'Semantic Parameter'," in Susan Rothstein, ed., *Events and Grammar* (Boston: Kluwer, 1998), pp. 53–103.

These representations all presuppose attentional capacities to select sets of individuals, and to keep track of which sets are being quantified. The input to analog magnitude representations are sets of dots, objects, sounds, actions, and for a representation of approximate cardinal value to do the infant or animal any good, it must be predicated of some particular set. Similarly, in parallel individuation, when the child compares two sets on the basis of 1–1 correspondence, she must be able to keep track of which is which. And natural language quantifiers explicitly quantify particular sets of individuals in various ways. However, as I now show, none of these quantificational resources deploy symbols for integers.

CONCEPTUAL DISCONTINUITY

In cases of conceptual discontinuity, no antecedent system of representation expresses the concepts of a later developing one. To establish discontinuity, then, one must characterize two empirically supported successive systems of representation (conceptual systems 1 and 2; CS1 and CS2), and demonstrate why CS2 had more expressive power than CS1. Let us consider each of the above described representational systems as CS1s. None of these systems of representations contains any symbols that represent any natural numbers.

The ontogenetically earliest system of representation that contains symbols for natural numbers (at least a finite subset of them), is the verbal numeral list representation of the positive integers, characterized by Gelman and Gallistel's[42] counting principles. The counting principles: the numeral list is strictly ordered (stable order); in counting, numerals are assigned to individuals in a set in 1–1 correspondence (1–1 correspondence); the last numeral assigned represents the cardinal value of the set (the cardinality principle). This system of representation is typically mastered by age 3 ½ or 4 by children growing up in a numerate culture. It expresses a finite subset of the natural numbers. The principles that underlie it are recognized to be extendable to arbitrarily large numbers by children a few years older,[43] at which point children have constructed a representation of natural number.

Just as was the case of the three systems of mental representation described above, one may ask how the integer list, when deployed in counting, represents number. The explicit symbols ("one," "two," and so on) represent exact cardinal values of sets. However, much of the numerical content of the numeral list is implicit. The counting principles guarantee that the known numerals represent natural numbers, because they guarantee that they respect the successor function (that is, "seven" means 1 more than 6, which is one more than 5, and so on). But there is obviously no explicit representation of the Peano-Dedekind axioms, and this system of representation when first mastered serves only to enumerate sets of individuals.

---

[42] Gelman and Gallistel, *The Child's Understanding of Number.*

[43] P. Hartnett and Gelman, "Early Understandings of Numbers: Paths or Barriers to the Construction of New Understandings?" *Learning and Instruction,* VIII, 4 (1998): 341–74.

Still, it is important to stress what an achievement the numeral list representation is, both historically and ontogenetically (see below). It is not cross-culturally universal, nor is it even the historically earliest external symbol system with numerical content (rather, tally systems were[44]). I will take it to be CS2. I now turn to the discontinuity between the innate symbol systems and this CS2, the integer list representational system.

Analog magnitude representations of number lack the expressive power of any system of representation of the natural numbers, even the numeral list representation of a finite subset of integers, in two crucial respects. First, because analog magnitude representations are inexact and discriminable only to a given Weber ratio, they fail to capture small numerical differences between large sets of objects. The distinction between eight and nine, for example, cannot be captured reliably by the analog magnitude representations of human adults. For any creature's analog magnitude system there will be a number, n, such that the distinction between n and $n + 1$ will not be captured reliably. Relatedly, analog magnitude representations of cardinal values are not built around, nor can they express, the successor function. Rather, they positively obscure the successor function. They contain no representation for exactly one. And since numerical values are compared by computing a ratio, the difference between one and two is experienced as different from that between two and three, which is again experienced as different from that between three and four. And, of course, the difference between twelve and thirteen is not experienced at all, since twelve and thirteen, like any higher successive numerical values, cannot be discriminated. In sum, analog magnitude representations are not powerful enough to represent the natural numbers. They do not provide exact representations of numbers and they obscure the successor function, which is constitutive of natural number.

---

[44] See James Hurford, *Language and Number: The Emergence of a Cognitive System* (New York: Blackwell, 1987).

The parallel individuation system does not remotely have the capacity to represent natural number. Unlike the analog magnitude number representation system, the parallel individuation system is not dedicated to number representations. Number is only implicitly represented, in that computations of 1–1 correspondence are made over symbols for individuals represented in parallel in models of arrays of objects and events. The system of parallel individuation contains machinery for indexing and tracking sets of individuals, but it contains no symbols for cardinal values. The only symbols in such models represent the individuals themselves. Also, the system of parallel individuation has an upper bound at very low set sizes indeed—three for infants. With this system of representation, infants cannot even represent 4 (even implicitly), let alone 7 or 32 or 1,345,698.

Verbal numerals are quantifiers, but the universal system of natural language quantifiers (words or morphemes that express concepts like *singular, plural, all, each, some,* and so on) do not represent integers. They do not express exact cardinal values of sets.

In sum, none of the three innate systems of representation with numerical content has the capacity to represent even a finite subset of the natural numbers, whereas CS2 (the numeral list representation of number) does. CS2 is discontinuous with any attested CS1.

Empirical evidence for conceptual discontinuity. Hypothesized conceptual discontinuities make three crucial predictions. First, CS2 should be very difficult to learn. Second, learners should make systematic, far-reaching errors, as they initially assimilate the evidence for CS2 to representational systems very different from it. Both of these predictions are confirmed in the case of children's mastering the numeral list representation of natural number. Third, CS2 should not be a feature of all cultures in the historical record, whereas the innate representational systems should be. This third prediction is also borne out in ethnographic studies of numerical cognition in adults from cultures with no numerical list, as well as studies of the historical construction of this representational resource.[45]

Children learn to count during the ages of 2 to 4 years, and learning to count is far from easy. Although young toddlers use a stably ordered list and count each object just once, honoring the stable order and 1–1 correspondence principles (two of the counting principles that guarantee that counting represents natural number), they do so for almost a year and a half before they figure out the cardinality principle (that the last number reached in a count represents the cardinal value of the set)—that is, before they figure out how counting represents number.[46] Initially, even after they have learned to recite 6 to 10 verbal numerals in order, children assign no numerical meaning to any of these number words, not even "one." They cannot even hand a person "one penny" (from several) if asked, or indicate which of two sets (a set of one and a set of several) has "one."

---

[45] For example, see P. Gordon, "Numerical Cognition without Words: Evidence from Amazonia," *Science*, cccvi, 5695 (2004): 496–99; Hurford, *Language and Number*; P. Pica, C. Lerner, and V. Izard, "Exact and Approximate Arithmetic in an Amazonian Indigene Group," *Science*, cccvi, 5695 (2004): 499–503.

Karen Wynn[47] studied very young counters—children who had a stably ordered count list to "ten" and who could apply the count routine to a set of objects, respecting 1–1 correspondence. She showed that from very early in the process of learning to count, children work out the meaning of "one." They can give you one penny from a pile of pennies if asked to, and they correctly distinguish a card with one fish from a card with any other number of fish if asked to indicate the card with "one fish." But they have not assigned a cardinal meaning to any other numeral on their count list. The striking phenomenon Wynn discovered (and you can definitely try this at home if you have a handy two-year-old) is that if you ask for "two pennies" or "four pennies," many young children merely grab a handful, with no relations to the number requested. That is, they do not give more when asked for four than when asked for two. Notice, they must know that the other words in the count sequence contrast with "one." They always grab a random number of objects greater than one when asked to hand over "two, three, four ..." objects. If asked to indicate the card with "three" fish, they successfully point to a card with three fish when it is contrasted with a card with one, even though their choices are random when they are offered a choice between cards with three and two. Such children are called "one"-knowers, for they know the meaning only of the verbal numeral "one." They have no idea which particular cardinality any of their other numerals refer to.

Consistent with the claims that preverbal representations of number are quite different from counting, and that learning the meaning of the count list is very difficult, children remain "one"-knowers for 6 to 8 months before they figure out what "two" means. Upon becoming

---

[46] Karen C. Fuson, *Children's Counting and Concepts of Number* (New York: Springer, 1998); M. LeCorre, E.M. Brannon, G.A. Van de Walle, and Carey, "Re-visiting the Competence/Performance Debate in the Acquisition of the Counting Principles," *Cognitive Psychology*, LII, 2 (2006): 130–69; Wynn, "Children's Understanding of Counting," *Cognition*, XXXVI (1990): 155–93, and "Children's Acquisition of the Number Words and the Counting System," *Cognitive Psychology*, XXIV (1992): 220–51.

[47] Wynn, "Children's Understanding of Counting," and "Children's Acquisition of Number Words and the Counting System."

"two"-knowers, they can construct sets of two when asked for "two," and tell which of two cards has "two fish" when given a contrast between an array of 2 fish and an array of 1 or an array of 3 or more. But they still respond randomly on "three." They are "two"-knowers for several months, and then become "three"-knowers. LeCorre and Carey[48] showed that all children who have worked out how counting represents number have previously assigned "one" through "four" numerical meanings. Thus, children laboriously work out numerical meanings for the numerals "one" through "four" over a 1 and 1/2 year period before they figure out how to use counting to implement the successor function. Children who have not yet figured out the cardinality principle (the last numeral reached in a count represents the cardinal value of the set) are called "subset-knowers," for they have assigned cardinal meanings to only a subset of the numerals in the count list. Those who have worked out how counting represents number are called "cardinal principle-knowers."

The ways in which CS2 transcends core cognition makes sense of why it takes children a year and a half or two years to figure out how counting represents number. Not only does the discontinuity hypothesis require that constructing CS2 should be difficult, but it also predicts that behavioral measures will reflect a qualitative change in representational capacity as CS2 is constructed. We should see within-child consistency on a whole variety of tasks that reflect CS2. And indeed we do. There is evidence from a wide variety of measures for a qualitative shift between subset-knowers, on the one hand, and cardinal principle-knowers, on the other. There is also evidence for consistency within knower-level. A "one"-knower reveals knowledge only of the numeral "one" on every task that probes for such knowledge. Ditto for "two"-, "three"-, and "four"-knowers.

LeCorre, Brannon, Van de Walle, and Carey[49] provide a thorough documentation of these claims; here I give just a few examples of the striking within-child consistency on several measures that suggests a qualitative shift in understanding how counting represents number upon becoming a cardinal principle-knower. In Wynn's original studies,[50] subset-knowers almost never counted to produce sets (asked to give five apples, they merely grabbed a handful), whereas cardinal principle-knowers almost always counted out large sets. Also, when simply asked to count a set of objects, children in both groups could do so with few errors, but then after counting, if asked "How many was that?" the cardinal principle-knowers almost always merely repeated the last word of their previous count, whereas subset-knowers rarely did so. Rather, subset-knowers recount, or very often provide a numeral that does not match the last word of their count. This suggests that subset-knowers do not realize that the last word reached in a count represents the cardinal value of the set. Later studies confirmed these findings and extended them. Even cardinal principle-knowers sometimes make mistakes when creating sets of a requested number. Children are asked to count and check their answers. When the count reveals an incorrect set-size, cardinal principle knowers virtually always correct appropriately. Subset-knowers, in contrast, leave the set unchanged or correct in the wrong direction (for example, add more objects when the count revealed that there were already too many) on more than 70% of the trials.[51] These analyses show that subset-knowers differ qualitatively from cardinal principle-knowers with respect to the representations that support the meanings they have assigned verbal numerals.

---

[48] LeCorre and Carey, "One, Two, Three, Four, Nothing More: How Numerals Are Mapped Onto Core Knowledge of Number In the Acquisition of the Counting Principles," *Cognition*, cv (2007): 395–438.

[49] LeCorre, Brannon, Van de Walle, and Carey, "Re-visiting the Competence/Performance Debate in the Acquisition of the Counting Principles."

[50] Wynn, "Children's Understanding of Counting," and "Children's Acquisition of the Number Words and the Counting System."

Within-child consistency is also found within knower levels. Being a "one"-knower according to Wynn's "give-a-number" task predicts which pairs of sets children will succeed on when asked "which is *n*" (that is, any contrast between a set of one and a set with any other number; no other contrast). Ditto for "two"-, "three"-, and "four"-knowers. Similarly, knower level on Wynn's task predicts which set sizes children can successfully estimate (without counting) when simply shown a set of entities and asked how many it contains.[52] A "two"-knower says "one" for sets of 1 and says "two" for sets of 2 and uses higher numerals randomly for sets from 3 to 10.

In sum, two kinds of analyses support the claim that the representational system that is the verbal count list is discontinuous with antecedent representations. Most importantly, I have offered empirically supported characterizations of three such antecedent systems of representations with numerical content, and shown how they lack the expressive power of the count list. Second, I have offered evidence that the meanings of verbal numerals are, as predicted, very difficult to learn, and are systematically misinterpreted during the learning process.

---

[51] Wynn, *ibid.*; LeCorre et al., "Re-visiting the Competence/Performance Debate in the Acquisition of the Counting Principles."

[52] Le Corre et al., *ibid.*; LeCorre and Carey, "One, Two, Three, Four, Nothing More."

THE EXPLANATORY CHALLENGE: QUINIAN BOOTSTRAPPING

I turn now to the explanatory challenge: what learning processes can create representational resources with more expressive power than, or qualitatively different from, their input? This is one place where the project of characterizing the logical toolbox available to human beings, including representational capacities drawn upon in number representations, is important. In the bootstrapping mechanism sketched below, I appeal to an old favorite—the uniquely human capacity to create external symbols. I assume that the capacity for human language—the capacity to form lexical, syntactic, and morphological representations—has innate support, as do the logical capacities needed to express meanings and the inferential relations among propositions, and to engage in a variety of modeling processes. I will help myself to many of these capacities in accounting for the capacity to represent natural number. But the human capacities for language and logic do not, by themselves, account for the origin of concepts like *32*, *infinity*, *gene*, *cancer*, and *democracy*. We seek specific proposals for how new conceptual systems (in this case a system of natural number representations) are built.

This explanatory challenge of accounting for specific conceptual innovations has been extensively discussed by historians and philosophers of science, and many appeal to what are called "bootstrapping" processes as an explanation for how representational resources that transcend their input can be created. The very word "bootstrapping" is a hyperbolic metaphor, meant to capture the deep difficulty of the problem. After all, it is impossible to pull oneself up by one's bootstraps. Neurath's metaphor of building a boat while already in the middle of the ocean also captures the difficulty of the problem—that while not grounded one must build a structure that will float and support you. Not grounded in this case means that the planks one is building the boat with are not interpreted concepts one already represents. In other metaphors, the learner's concepts are partially grounded, as in Quine's ladder metaphor. Here, one builds a ladder grounded in one conceptual system until one has a platform that is self-sustaining, and then one kicks the ladder out from under. And in a final Quinian metaphor, one is scrambling up a chimney supporting oneself by pressing against the sides one is building as one goes along. Quine's metaphor captures the fact that new conceptual systems are used as they are built, even if they do not yet have their final form. This metaphor stresses, as does Neurath's boat, that the structure one builds consists of relations among the concepts one will eventually attain—it is that structure of interrelations among the to-be-attained concepts (the sides of the evolving chimney, the boat itself, the platform from which the ladder can be kicked away) that serves the crucial bootstrapping role. See Quine[53] for his own elaboration of Quinian bootstrapping.

Although such metaphors are evocative—of both the problem to be solved and the solution—Quine never describes in detail how this learning process operates. These metaphors are hardly satisfying to a cognitive scientist trying to understand bootstrapping mechanisms. As I illustrate here, I believe it is possible to flesh out the metaphors with appeals to processes that are fairly well understood at a computational level.

Quinian bootstrapping processes require explicit symbols, such as those in written and spoken language or mathematical notational systems. The aspect of the bootstrapping metaphor that consists of building a structure while not grounded is captured by the utterly unremarkable fact that learners often master a set of new terms together, and represent some interrelations among a new system of symbols to one another, directly, rather than by mapping each symbol onto preexisting concepts.[54] The symbols so represented thus serve as placeholders, at most only partially interpreted with respect to antecedent concepts. In historical cases, placeholder structures are often created by abductive leaps (for example, Maxwell's guess that the mathematics of Newtonian forces in fluid media would be the relevant mathematics to model Faraday's phenomena[55]). In ontogenetic cases placeholder structures are formed through ordinary language learning; a set of terms are acquired, interdefined, but not yet interpreted in terms of already existing representations. This is one essential component of Quinian bootstrapping.

The second essential component is the process through which the placeholders become interpreted. As historian and philosopher of science Nancy Nersessian argues (*ibid.*), these are modeling processes. Often, but not always, processes of analogical mapping are involved. Other modeling processes, such as thought experimentation and limiting case analyses, have roles in Quinian bootstrapping as well. Several properties of these modeling processes are important. They are not deductive, and there are no guarantees in bootstrapping. The structures that are tentatively posited either work, in the sense of continuing to capture the observed data that constrain them, or they do not. Finally, the modeling processes are abstraction processes—they serve to combine distinct representational resources, and to make explicit what is implicit. They are sensitive to the constraints that are implicit in each system, seeking to honor them wherever possible.

---

[53] Quine, *Word and Object*, *Ontological Relativity*, and *Roots of Reference*.

[54] Ned Block, "Advertisement for a Semantics for Psychology," in Peter A. French, ed., *Midwest Studies in Philosophy*, Volume 11, *Studies in Essentialism* (Minneapolis: Minnesota UP, 1986), pp. 615–78.

[55] Nersessian, "How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science," in Ronald Giere, ed., *Cognitive Models of Science* (Minneapolis: Minnesota UP, 1992), pp. 3–44.

Bootstrapping the numeral list representation of natural number. The output of the learning process of interest here is the numeral list representation of natural number. The problem of how the child builds a numeral list representation decomposes into the related subproblems of learning the ordered list itself ("one, two, three, four, five, six ..."), learning the meaning of each symbol on the list (for example, "three" means three and "seven" means seven), and learning how the list itself represents number, such that the child can infer the meaning of a newly mastered numeral symbol (for example, "eleven") from its position in the numeral list.

The child first learns "one, two, three, four, five ..." as a list of meaningless lexical items. There is no doubt that children have the capacity to learn meaningless ordered lists of words—they learn sequences such as "eeny, meeny, miny, mo," the alphabet, the days of the week, and so on. Indeed, nonhuman primates have this capacity; it is part of innate computational machinery.[56] This step in the learning process—learning an arbitrary ordered list ("one, two, three, four, five, six ...") is a paradigmatic example of one aspect of Quinian bootstrapping: the meanings of the counting words are exhausted, initially, by their interrelations, their relative order in the list, and their place in the numerically meaningless count routine. At this point in the process, the verbal numerals are placeholders with respect to the numerical meaning they will come to have.

Children do not learn the numerical meaning of "one" in the context of counting. "One" is much more frequent in speech to children as a quantifier than embedded in the count routine.[57] "Can you give me one?" "Would you like that one?" "I'd like one cupcake." As Paul Bloom and Wynn[58] demonstrated, from the outset of numeral production, children correctly use numerals in the syntactic positions of quantifiers. Bloom and Wynn conjectured that the semantics of quantifiers helps children recognize the numerical meaning of "one." Three sources of evidence support Bloom and Wynn's conjecture. First of all, the partial meanings children in the subset-knower stage assign to numerals implicate hypothesis testing over the space of quantifier meanings. For example, for "one"-knowers, "two, three, four," and so on each mean essentially "plural" or "some." Second, learners of Japanese and Mandarin, classifier languages with no singular-plural distinction, learn to count as early as do learners of English, and have equivalent number word input, but do not become "one"-knowers until several months later than do English-speakers.[59] Relatively sparse number marking in syntax slows down assigning partial meanings to the placeholder symbols in the count list. Finally, Palestinian Arabic has a dual marker system and also distinguishes plural morphology and collective morphology. In a study of children's learning number marking in Palestinian Arabic, D. Ravid and L. Hayek[60] found that 3-year-olds often used the numeral translated "two" instead of the dual when referring to sets of two objects, whereas older children were unlikely to do this. This finding is consistent with the suggestion that "two" is initially a dual marker. Dual markers are part of the machinery of natural language quantification that distinguish pairs of individuals from other pluralities.

---

[56] For example, Terrace, Son, and Brannon.

[57] B.W. Sarnecka, V.G. Kamenskaya, Y. Yamana, T. Ogura, and J.B. Yudovina, "From Grammatical Number to Exact Numbers: Early Meanings of 'One,' 'Two,' and 'Three,' in English, Russian, and Japanese," *Cognitive Psychology*, LV, 2 (2007): 136–68.

[58] Bloom and Wynn, "Linguistic Cues in the Acquisition of Number Words," *Journal of Child Language*, XXIV, 3 (1997): 511–33.

Before we can understand how children work out the meaning of the count list, we must establish what "one," "two," "three," and "four" mean for a "four"-knower, for it is only after being "four"-knowers that children figure out the cardinal principle. What is the format of the mental representations that underlie the numerical meanings subset-knowers have created for numerals? If "two" is a dual marker for "two"-knowers, what representations give numerical meaning to dual markers? What is the process through which a given set is assigned one numeral rather than another?

A system of representations that draw both on the resources of natural language quantification and parallel individuation could underlie the meanings of numerals for subset-knowers. LeCorre and I dubbed the proposed system "enriched parallel individuation."[61]

The innate parallel individuation system creates working-memory models of sets. The symbols in these models represent particular individuals—this box, which is different from that one. However, as detailed above, even when drawing on parallel individuation alone, infants have the capacity to represent two models and compare them on the basis of 1–1 correspondence. For representations of this format to subserve the meanings of the singular determiner or the numeral "one" for subset-knowers, the child may create a long-term memory model of a set of one individual and map it to the linguistic expression "a" or "one." Similarly, a long-term memory model of a set of two individuals could be created and mapped to the linguistic expression for a dual marker or "two," and so on for "three" and "four." These models could contain abstract symbols for individuals ({*i*}, {*j k*}, {*m n o*}, {*w x y z*}) or they could simply be long-term memory models of particular sets of individuals ({my head}, {my hand, my hand}...). What makes these models represent "one," "two," and so forth is their computational role. They are deployed in assigning numerals to sets as follows: The child makes a working-memory model of a particular set he or she wants to quantify, for example {cookie cookie}. He or she then searches the models in long-term memory to find that which can be put in 1–1 correspondence with this working-memory model, retrieving the quantifier that has been mapped to that model.

---

[59] Sarnecka et al., "From Grammatical Number to Exact Numbers"; P. Li, Le Corre, R. Shui, G. Jia, and Carey, "Effects of Plural Syntax on Number Word Learning: A Cross-linguistic Study," presented at the 28th Boston University Conference on Language Development (2003).

[60] Ravid and Hayek, "Learning about Different Ways of Expressing Number in the Development of Palestinian Arabic," *First Language*, XXIII, 1 (2003): 41–63.

[61] LeCorre and Carey, "One, Two, Three, Four, Nothing More"; see Kelly Mix, Janellen Huttenlocher, and Susan Cohen Levine, *Quantitative Development in Infancy and Early Childhood* (New York: Oxford, 2002) for a similar proposal.

This proposal is motivated by the observation that the subset-knower stage encompasses assigning meanings to the numerals "one" through "four" only, for children are limited to working memory models of four individuals represented in parallel. It is also consistent with the results of an unpublished training study by Y-Ting Huang. Huang taught "two"-knowers to reliably identify three butterflies as "three," when faced with choices of 3 versus 2 and with choices of 3 versus 4 or more. That is, she attempted to turn a "two"-knower into a "three"-knower. Children could learn to do this, and they generalized, "three" to triples of butterflies they had not observed in training, butterflies of vastly different sizes, shapes and markings. However, they failed to apply "three" to sets of balls, or cats, or chairs. And they were still "two"-knowers according to Wynn's Give-a-Number task. Apparently, the training led them to establish a model of a triple of butterflies, and they could compare working memory models of other arrays of butterflies to this model on the basis of 1–1 correspondence, but it did not occur to them that any array that could be put in 1–1 correspondence with this model should be called "three."

All of the computational resources required for enriched parallel individuation are known to be available to prelinguistic infants.[62] Prelinguistic infants create working-memory models of at least two separate sets (recall the cracker choice experiments described above) and compare these on the basis of 1–1 correspondence (recall the search-into-the-box experiments described above). They also treat sets as objects, quantifying over them as required by natural language quantifiers. Still, it is important to stress that the long-term memory models that support the meanings of singular, dual, and triple markers, as well as the child's first numerals, are not themselves part of the innate representational repetoire. These must be created in the course of language learning, and for English-learning children this process unfolds for a period of over a year.

---

[62] See above, and Carey, *The Origins of Concepts,* for full evidence for this claim.

The two important planks of the bootstrapping process are constructed in parallel, largely independent of each other. First, the child learns the explicit numeral list together with the count routine as a numerically meaningless game. Second, the child creates numerical meanings for some numerals—in this case "one" through "four." These meanings are supplied by enriched parallel individuation, and so far, no bootstrapping has occurred.

The stage is set for the completion of the bootstrapping processes. Children use the placeholder structure to model sets of individuals in the world. They note the identity of the words "one," "two," "three," and "four" which now have numerical meaning, and the first words in the otherwise meaningless counting list. Also, in the course of counting, children discover that when an attended set would be quantified with the dual marker "two," the count goes "one, two," and when an attended set would be quantified with the trial marker "three," the count goes "one, two, three." The child is thus in the position to notice that for these words at least, the last word reached in a count refers to the cardinal value of the whole set.

At this point, the child makes the crucial induction. The child must notice an analogy between next in the numeral list and next in the series of mental models ($\{i\}$, $\{j\ k\}$, $\{m\ n\ o\}$, $\{w\ x\ y\ z\}$) related by adding an individual. Remember, core cognition supports the comparison of two sets simultaneously held in memory on the basis of 1–1 correspondence, so the child has the capacity to represent this latter basis of ordering sets. This analogy licenses the crucial induction: if "x" is followed by "y" in the counting sequence, adding an individual to a set with cardinal value x results in a set with cardinal value y. This generalization does not yet embody the arithmetic successor function, but one additional step is all that is needed. Since the child has already mapped single individuals onto "one," adding a new individual is equivalent to adding one.

This proposal makes sense of the actual partial meanings children assign to number words as they try to fill in the placeholders. The semantics of quantifiers explain these facts. It makes sense of the fact that subset-knowers acquire the cardinal meanings of "one," "two," "three," and "four," and no other numeral, for only sets of these sizes are representable by models of the sets of individuals held in parallel in working memory, thus to be matched via 1–1 correspondence to long-term memory models of sets of one, two, three, and four individuals. Finally, Sarnecka and Carey[63] provided direct evidence that the difference between subset-knowers and cardinal principle-knowers is a (tacit) appreciation of how counting implements the successor function, precisely the induction posited in the above bootstrapping proposal. Cardinal principle-knowers, but not subset-knowers, judged that if a hidden array they were told consisted of 5 objects had another added to it, the resultant array must have 6 not 4 or 7.

We sought an answer to several questions. How do children assign numerical meanings to verbal numerals, and how do children learn how the list itself represents number? The bootstrapping proposal provides answers to both. The meanings of "one" through "four" are acquired just as quantifiers in natural languages are—as quantifiers for single individuals, pairs, triples, and quadruples. These words, as well as higher numerals, also get initial interpretations as part of a placeholder structure, the count list itself, in which meaning is exhausted by the fact that the list is ordered. The bootstrapping process explains how children learn how the list itself represents number, which in turn explains how they assign numerical meaning to numerals like "five" and "seven." When children first become cardinal principle-knowers, the meaning of "five" is exhausted by the child's mastery of counting. The counting principles ensure that the content of "five" is one more than four, and the meaning of "seven" is one more than six, which is one more than five, which is one more than four, and so on.

Interim conclusions—the construction of representations of natural number. One surprising upshot of the above proposal is that one of the evolutionarily ancient systems of representation with numerical content, the analog magnitude system, plays no role in providing initial meanings for verbal numerals. LeCorre and Carey[64] provide extensive evidence that this is so. However, a further bootstrapping episode, within months of working how the count-list represents number, integrates the numeral list with analog magnitude number representation, greatly enriching their numerical content.

---

[63] Sarnecka and Carey, "How Counting Represents Number: What Children Must Learn and When They Learn It," *Cognition*, CVIII, 3 (2008): 662–74.

[64] LeCorre and Carey, "One, Two, Three, Four, Nothing More."

Clearly, when children have learned to use a count list that extends to 10 or 20 to enumerate sets, they have not created a representation of natural number. They may not have yet abstracted a concept *number* at all. However, children spontaneously use this new representational resource to invent algorithms for arithmetic (for example, adding by counting on), and by age 5 or 6 many spontaneously invent arguments that there is no highest number (for example, one five-year old said to me: "suppose you think a gazillion is the highest number—well, you can go a gazillion and one, a gazillion and two ..."). The integer list, as part of the count routine, is an essential representational structure along one route to constructing representations of integers.

I have argued here that the numeral list representation of number is a representational resource with power that transcends any single representational system available to prelinguistic infants. When the child, at around age 3 ½, has mastered how the count sequence represents number, he or she can represent any exact cardinality expressed in their count list. Before that, he or she has only the quantificational resources of natural languages, parallel individuation representations that implicitly represent small numbers, and analog magnitude representations that provide approximate representations of the cardinal values of sets.

Additionally, I have taken on the challenge of specifying a learning mechanism that can underlie specific developmental discontinuities—Quinian bootstrapping. Quinian bootstrapping involves, but is not exhausted by, garden-variety learning processes: association, the mechanisms that support language learning, and so on. In addition it involves noticing analogies and making inductive and abductive leaps. The specific bootstrapping proposal sketched here depends on the analogy between next on the numeral list and next state after additional individual has been added to a set.

In Quinian bootstrapping, an explicit structure is learned initially without the meaning it will eventually have, and at least some relations among the explicit symbols are learned directly in terms of each other. The list of numeral words and the counting routine are learned as numerically meaningless structures. Whereas order is essential to numerical representations, ordered relations in themselves are much more general and thus not uniquely numerical. The ordering of the number words exhausts their initial representational content within the counting routine and plays a role in the mappings and inductions through which counting comes to have numerical content.

Quinian bootstrapping depends upon integrating previously distinct representations. This is where the new representational power comes from. The concepts *set, individual, singular, plural,* and perhaps *dual* and *triple* are explicitly available to support the learning of quantifiers, but are only implicit or absent in parallel individuation and analog magnitude representations. In analog magnitude representations, numerical distinctions are explicitly symbolized that are unmarked in natural language quantifiers or parallel individuation (for example, 35 versus 40), and although analog magnitude representations may play no role in the child's learning how counting represents number, they are integrated with counting some six months later.[65] And the representations that articulate the parallel individuation system contain computations that embody the successor function, whereas neither of the other systems does. The bootstrapping process (which depends on analogical mapping) creates an explicit representational system with all of these properties, a representational system that maps onto each of its sources and thus serves to integrate them.

Kronecker was wrong. Natural number is a human construction. I have provided here one worked example of the creation of a new representational resource with more power than the representations upon which it is built. The lessons I wish to draw, however, are very general. Such creations occur repeatedly, both historically and within the individual child. Within mathematical representations, much has been written about the creation of 0, negative numbers, and rational numbers. Each of these developments transcends the power of the numeral list representation that is the focus of the present paper, and each requires further episodes of Quinian bootstrapping.[66] Similarly, within the history of science, theory changes that involve conceptual change require the creation of new representational resources that allow thoughts that were previously unthinkable, and these theory changes also require Quinian bootstrapping.[67]

SOME FINAL LESSONS

In *The Origins of Concepts*, I draw several general lessons from case studies such as that sketched here. Concerning the existence and nature of innate conceptual representations, the analog magnitude system of representation is one of several systems of core cognition. Systems of core cognition include innate representations with conceptual content. They are supported by innate perceptual input analyzers that identify the entities in their domains, they have long evolutionary histories, and they operate continuously throughout the life span. The format of representation in core cognition is iconic. It is an empirical claim that there are systems of representation with these properties. Analog magnitude number representations do. Another system of core cognition includes representations of objects and their physical interactions and yet another includes representations of agents, their goals, and other intentional states.

---

[65] LeCorre and Carey, *ibid.*

[66] See Carey, *The Origins of Concepts*, for a discussion of the construction of rational number in middle childhood.

[67] See Carey, *The Origins of Concepts*.

Second, most of human cognition does not have the properties of core cognition: it is not innate, it is not shared with other animals, its symbols are not iconic nor the output of innate perceptual input analyzers, and it is not continuous throughout development. Thus, a theory of conceptual development must account for the human capacity for conceptual systems so utterly different from those shared with other animals. Clearly, the human language faculty is at the heart of such a theory. Beyond garden variety language learning, the human capacity for external symbols is essential for the creation of the placeholder structures of Quinian bootstrapping which in turn play a central role in conceptual discontinuity.

Finally, a full account of particular representational systems (origin, format, and computational role) allows us to see what determines the content of the symbols therein. With respect to core cognition, an information semantics applies—the innate perceptual mechanisms that identify entities in the domain and innate conceptual role guarantee that the symbols are causally connected to those entities they represent. With respect to symbols in later developing representational systems, both information semantics and conceptual role semantics have parts to play in content determination. The whole story includes some speculations about how the details of the acquisition process might help solve the problem of characterizing which aspects of conceptual role are content determining. For the whole story, see *The Origins of Concepts.*

SUSAN CAREY

Harvard University

THE JOURNAL OF PHILOSOPHY

Special Issues Available

Epistemic Norms - September and October 2008
Parts and Wholes - December 2006
On Bernard Williams - June 2003
Causation - April 2000
Jürgen Habermas and John Rawls - March 1995

The Dewey Lectures:
Michael Dummett - *Truth and the Past,* 2003
Hilary Putnam - *Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind,* 1994
Donald Davidson - *The Structure and Nature of Truth,* 1990
Amartya Sen - *Well-being, Agency, and Freedom,* 1985
John Rawls - *Kantian Constructivism in Moral Theory,* 1980
Ernest Nagel - *Teleology Revisited,* 1977
W.V. Quine - *Ontological Relativity,* 1968

The Woodbridge Lectures:
John McDowell - *Having the World in View: Sellers, Kant, and Intentionality,* 1997

Subscriptions

Individuals $45.00
Students $20.00
Libraries and Institutions $100.00
Print and Online for Libraries and Institutions $180.00
Postage outside the United States $20.00

www.journalofphilosophy.org

The Monist

An International Journal of General Philosophical Inquiry

JANUARY 2009 ◆ VOL. 92, NO. 1 ◆ SINGLE ISSUE, US $12

SINGULAR CAUSATION

ADVISORY EDITOR: MICHAEL MOORE

ARTICLES

MICHAEL MOORE	*Introducton: The Nature of Singularist Theories of Causation*
PHIL DOWE	*Absences, Possible Causation, and the Problem of Non-Locality*
DOUGLAS EHRING	*Abstracting Away from Preemption*
CHRISTOPHER HITCHCOCK	*Problems for the Conserved Quantity Theory: Counterexamples, Circularity, and Redundancy*
STEPHEN MUMFORD	*Passing Powers Around*
BENCE NANAY	*The Properties of Singular Causation*
MICHAEL ROTA	*An Anti-Reductionist Account of Singular Causation*
JESSICA WILSON	*Resemblance-based Resources for Reductive Singularism*

❧

Founded in 1888 by Edward C. Hegeler, *The Monist* is one of the oldest and most important philosophical journals in the world. Established originally as a quarterly journal of the philosophy of science, *The Monist* was edited by Paul Carus until his death in 1919. Currently edited by Barry Smith, each issue addresses a particular philosophical topic and stands as a collected anthology of continuing interest.

Editor: Barry Smith, University at Buffalo, State University of New York

Managing Editor: George A. Reisch (new managing editor): themonist@gmail.com

Single issues, $12 postpaid. Annual subscription: institutions, US $55; individuals, US $35; make checks payable to THE MONIST, 315 Fifth Street, Peru, IL U.S.A. 61354.

Please see our Web site for foreign subscription information: monist.buffalo.edu

CAMBRIDGE

New and Noteworthy Titles

The Cambridge Companion to Galen
Edited by R. J. Hankinson
Cambridge Companions to Philosophy
$85.00: Hb: 978-0-521-81954-1
$29.99: Pb: 978-0-521-52558-9:
476 pp.

The Cambridge Companion to Darwin 2nd Edition
Edited by Jonathan Hodge and Gregory Radick
Cambridge Companions to Philosophy
$90.00: Hb: 978-0-521-88475-4
$34.99. Pb: 978-0-521-71184-5:
562 pp.

Land, Conflict, and Justice
A Political Theory of Territory
Avery Kolers
$99.00: Hb: 978-0-521-51677-8:
252 pp.

Social Rights Jurisprudence
Emerging Trends in International and Comparative Law
Edited by Malcolm Langford
$160.00: Hb: 978-0-521-86094-9
$90.00: Pb: 978-0-521-67805-6:
704 pp.

A Cultural Theory of International Relations
Richard Ned Lebow
$120.00: Hb· 978-0-521-87136-5
$39.99: Pb: 978-0-521-69188-8:
774 pp.

The Cambridge Companion to Paul Tillich
Edited by Russell Re Manning
Cambridge Companions to Religion
$90.00: Hb: 978-0-521-85989-9
$29.99: Pb: 978-0-521-67735-6:
348 pp.

The Cambridge History of Jewish Philosophy
From Antiquity through the Seventeenth Century
Volume 1
Edited by Steven Nadler and T.M. Rudavsky
Cambridge History of Jewish Philosophy
$180.00: Hb: 978-0-521-84323-2:
916 pp.

Political Thought and History
Essays on Theory and Method
J.G.A. Pocock
$90.00: Hb: 978-0-521-88657-4
$28.99: Pb. 978-0-521-71406-8:
296 pp.

Paradoxes 3rd Edition
R. M. Sainsbury
$90.00. Hb: 978-0-521-89632-0
$29.99: Pb: 978-0-521-72079-3:
190 pp

Mill's *On Liberty*
A Critical Guide
Edited by C. L. Ten
Cambridge Critical Guides
$99.00: Hb. 978-0-521-87356-7:
252 pp

www.cambridge.org/us
CAMBRIDGE
UNIVERSITY PRESS
1584 · 2009
425 YEARS OF CAMBRIDGE PRINTING AND PUBLISHING

COLUMBIA

Learn more at www.cup.columbia.edu

Mind in the Balance

Meditation in Science, Buddhism, and Christianity

B. Alan Wallace

Our best-selling mind and consciousness scholar boldly corrects the balance between empirical study and religion.

"A significant contribution to the field."
—Robert Mayer, University of Oxford

cloth - $24.95
Columbia Series in Science and Religion

Strange Wonder

The Closure of Metaphysics and the Opening of Awe

Mary-Jane Rubenstein

"A genuinely original contribution to burning issues in contemporary philosophy, theology, and philosophical theology."
—Denys Turner, Yale University

cloth - $45.00
Insurrections: Critical Studies in Religion, Politics, and Culture

Self, Logic, and Figurative Thinking

Harwood Fisher

"A journey through difficult and perilous concepts—awareness, consciousness, perception—to engage the idea of the self as the original locus of consciousness."
—Luis Radford, Laurentian University

cloth - $40.00

NEW FROM EDINBURGH UNIVERSITY PRESS

Understanding Ethics

An Introduction to Moral Theory
Second Edition

Torbjörn Tännsjö

"An excellent book."—Jonathan Glover

"Clear, comprehensive, and fair-minded coverage of the field."—L. Wayne Sumner, *Theoria*

paper - $27.50 / cloth - $95.00

CENTRAL LIBRARY
14-8-09
UNIVERSITY OF CALCUTTA
CORE CAMPUS

THE JOURNAL OF PHILOSOPHY

VOLUME CVI, NUMBER 5
MAY 2009

page
255 *Thisness and Events* Joseph Diekemper
277 *Causal Independence, the Identity of Indiscernibles, and the Essentiality of Origins* Charles B. Cross

COMMENTS AND CRITICISM
292 *Collective Intentions and Game Theory* Raimo Tuomela

BOOK REVIEWS
301 Kathrin Koslicki: *The Structure of Objects* Trenton Merricks
308 MEMO TO AUTHORS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Michael J. Mooney; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2009

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.
POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2009 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 5, MAY 2009

UNIVERSITY OF CALCUTTA

THISNESS AND EVENTS*

This essay is an investigation into the existence of a very unusual and—some would say—unacceptably exotic type of property: namely, the property of being a certain individual; or, if you prefer, the property of being identical to a certain individual. In other words, this essay will investigate whether—in spite of their exotic nature—there are *thisnesses*, and, in particular, whether thisnesses are instantiated by events. Of course, I have not really said enough yet about thisnesses to motivate the criticism that such properties are exotic; for it might be the case that an individual *x*'s thisness is merely a conjunction of all of *x*'s qualitative properties. And, if we allow for the existence of all of the properties of an individual, it is not overly profligate to allow for the existence of this one additional conjunctive property. This, apparently, was Leibniz's view. Though seemingly ontologically mundane, however, such a qualitative[1] thisness is not ontologically innocent. Given the definition of a thisness above, an individual's thisness is essential to it, and there are few philosophers who would wish to hold that all of an individual's properties are essential to it.[2] Rather, it is the Scotistic[3] variety of thisness—according

---

* Earlier versions of this paper were presented at the University of St. Andrews and Queen's University Belfast. I am grateful to the audiences on those occasions for their helpful and stimulating comments. I am particularly grateful to Katherine Hawley and Cynthia Macdonald. The paper has also benefited considerably from comments supplied by the Editors.

[1] Assuming that a conjunction of basic qualitative properties is itself qualitative (though obviously not basic). See Robert Adams, "Primitive Thisness and Primitive Identity," this JOURNAL, LXXVI, 1 (January 1979): 5–26.

[2] Adams, *ibid.*, p. 10, notes that this unpalatable ramification of Leibniz's view can be avoided, without rejecting his conception of thisnesses as qualitative, by claiming that they are disjunctions of basic qualitative properties. Such disjunctions would allow for an individual to instantiate different basic qualitative properties in different worlds.

[3] After Duns Scotus, who coined the Latin term '*haecceitas*', of which 'thisness' is a translation.

0022-362X/09/0605/255–76

to which thisness is primitive and purely nonqualitative—that I am interested in pursuing here. And it is precisely the nonqualitative aspect of Scotistic thisness that strikes many as being unacceptably exotic.

What is it for a property to be nonqualitative? The distinction between a qualitative property and a nonqualitative property is one that belongs to that family of philosophical distinctions which, though not admitting of analysis, can be made easily enough through the use of a loose definition and some intuitive examples. Here is the definition offered by Robert Adams:

> a property is purely qualitative—a suchness—if and only if it could be expressed, in a language sufficiently rich, without the aid of such referential devices as proper names, proper adjectives and verbs (such as 'Leibnizian' and 'pegasizes'), indexical expressions, and referential uses of definite descriptions (*ibid.*, p. 7).

So, living in a large city and being the son of a carpenter are qualitative properties, while living in New York and being the son of Henry are nonqualitative properties. The reason nonqualitative thisness is sometimes thought to be ontologically suspect is that, by definition, we are unable to describe such a property other than through the use of such devices as demonstratives and proper nouns. Thus, for example, if the event of the Battle of Waterloo has nonqualitative thisness, then there exists[4] a property that is the property of being that particular Battle, or of being identical to it; and this property exists over and above all the Battle's more mundane, qualitative properties and relations, such as being bloody, being such as to have occurred 39 years after the American Revolution, and being a battle in which a diminutive French general was decisively defeated.

Now, suppose we were able to list all such qualitative properties of the Battle of Waterloo. The opponent of thisness will want to know what further property that event could possibly instantiate. If all we are after is the concept of what it is for an individual to be the individual that it is, then surely such a concept could be captured by an exhaustive list of all that individual's qualitative properties. The intuitive thought behind this objection is that ontological parsimony (not to mention sobriety) dictates against the postulation of mysterious individual essences, when the concept that such essences are meant to explain can be equally well explained by nonmysterious qualitative properties. So we can conceive of what it is to be a particular individual without claiming that there is a corresponding property that is had essentially by that individual, and that, necessarily, no other individual has.

---

[4] Or existed, depending on whether you think a past individual's thisness continues to exist. See Adams, "Time and Thisness," *Midwest Studies in Philosophy*, XI (1986): 315–29.

That is a brief account of the objector's motivation. What of the proponent of thisness? What motivates him to violate—by the objector's lights—principles of ontological parsimony? One thought is that the existence of nonqualitative thisness provides a plausible argument for the transworld identity of individuals; and, since many modal actualists will want to endorse such an identity, they may wish to avail themselves of thisness as well.[5] Another thought, from the philosophy of time rather than modality, is that the existence of the thisnesses of past events, and the nonexistence of thisnesses of future events, can provide the framework for a growing universe temporal ontology (according to which past events exist and future events do not).[6] I will not say much more about either of these motivations, other than to note that it is the latter thought that has piqued my own interest in nonqualitative thisness. The task I have set myself here, however, is to investigate whether some independent argument—independent, that is, of one's motivation—for the existence of nonqualitative thisness might be forthcoming. Nevertheless, stating my motivation from the outset will explain why I am particularly interested in whether events instantiate nonqualitative thisness.[7] It also explains why a favorable outcome—from my perspective—would involve a demonstration of the existence of thisness that is independent of the existence of human beings; since I take it that the existence and nature of time is logically independent of the existence and nature of human beings.

I will begin my investigation in section I with a look at what is commonly taken to be the test for nonqualitative thisness, namely, the failure of application of the principle of the Identity of Indiscernibles for some class of individuals. The two classes that we will look at are those of material objects and events. In section II, I will discuss Ian Hacking's objection to the general project of seeking counterexamples to the Identity of Indiscernibles, and then consider a response due to Adams. Finally, in section III, I will offer my own response to

---

[5] See Adams, "Primitive Thisness and Primitive Identity," and "Actualism and Thisness," *Synthese*, LVII (1981): 3–42. As has no doubt become clear, I have Adams in mind as my proponent of nonqualitative thisness (at least at this stage).

[6] See Adams, "Time and Thisness."

[7] The term 'instantiate', and its related forms, strike me as slightly unfortunate in this context. Given that thisnesses are like tropes in that they are unique, unrepeatable properties, it is not quite proper to speak of their having instances. Nonetheless, I like the alternatives to 'instantiation' even less, so I will continue to use it and its related forms.

Hacking, and then go on to draw some conclusions with respect to thisness and events.[8]

I. A TEST FOR THISNESS

I.A. The Principle of the Identity of Indiscernibles. It is widely acknowledged that the appropriate test for the existence of thisness involves finding some class of suitably arranged individuals for which the principle of the Identity of Indiscernibles does not apply. As has also been widely acknowledged, however, it is crucial to the legitimacy of the test that the principle take the proper form (proper, that is, for the intended purpose in this context). In its general form, the principle states that any two individuals which share all of their properties are identical. But what kind of properties are we interested in in this context—that is, in the context of applying a test for the existence of Scotistic thisness? Suppose that objects x and y share all their properties. One of x's properties is being identical to x, so, *ex hypothesi*, y has the property of being identical to x. Thus, when the principle is interpreted as quantifying over not only qualitative properties, but also nonqualitative properties, it is trivially true and therefore incapable of yielding interesting metaphysical theses. Given this, it is necessary to strengthen the principle by restricting the domain of properties quantified over to that of qualitative properties alone. For if it can be shown that individuals x and y share all their qualitative properties, and yet are distinct individuals, then there must be some further, nonqualitative property that accounts for their distinctness (though we will look at an objection to this inference in section I.B.1, below).

Notice, too, that we need not strengthen (in this respect) the principle any further; say, by restricting the domain of properties quantified over to that of nonrelational, qualitative properties. To do so would certainly yield a metaphysically interesting thesis, since, according to it, individuals are differentiated solely by their intrinsic qualitative properties. But, based on the above inference pattern, if we could show that objects x and y share all their intrinsic qualitative properties, and yet are distinct individuals, then either they are differentiated by thisness *or* by their relational qualitative properties. In which case, we would not have conclusively demonstrated the necessity of the former. The lesson is that, in testing for thisness, we should utilize the weakest, nontrivial form of the principle of the Identity of Indiscernibles. Restricting the domain of properties quantified over to that of all qualitative properties, both relational and nonrelational, satisfies this desideratum.[9]

---

[8] For the remainder of the paper, I intend 'thisness' to be read 'nonqualitative thisness'.

[9] But see Gonzalez Rodriguez-Pereyra, "How Not to Trivialise the Identity of Indiscernibles," in P.F. Strawson and A. Chakrabarti, eds., *Concepts, Properties and Qualities*

So, for our purposes here, there is at least one adjustment that must be made to the general form of the principle. Are there any others? What about the principle's modal status? If we interpret the principle as stating a contingent truth, as opposed to a necessary truth, then even this modally weak form of the principle will be of metaphysical interest. As we have seen, it is by no means trivial to claim that individuals are differentiated by their qualitative properties alone, since such a claim disbars nonqualitative properties as differentiating features; and this nontrivial claim would certainly not be trivialized by restricting the domain of individuals quantified over to that of actual individuals. Nevertheless, when we consider that we have very little reason to believe that there are two actual, distinct individuals which share all their qualitative properties, it becomes clear that the counterexamples we will be looking for in the course of the investigation will have to involve possible individuals, and such individuals can only serve as counterexamples to the principle in its necessary form. On the basis of these considerations, then, we are in a position to formulate the principle of the Identity of Indiscernibles that will best suit our purposes in this context:

> (I/I) Necessarily, any two individuals that have all the same qualitative properties, both relational and nonrelational, are the same individual.

I.B. Counterexamples to the Principle. Now that we have the proper form of the principle (I/I) in view, we can consider what types of possible individuals might serve as counterexamples to it. We can think of such counterexamples as potentially providing (i) a general conclusion about the existence of thisness, as well as (ii) a specific conclusion about the existence and instantiation of thisness in a particular class of individuals. So, for example, if we can devise a counterexample to (I/I) that involves material objects, then we can infer from this counterexample both that, generally, there are thisnesses, as well as that, specifically, material objects have them. This point has important consequences, since it may be that only certain classes of individuals instantiate thisness, so we should not infer from successful counterexamples that all individuals instantiate thisness, nor should we infer from unsuccessful counterexamples that no individuals instantiate thisness. The best way forward, then, will be to devise potential counterexamples to (I/I) involving the classes of individuals that we are most interested

---

(Aldershott, UK: Ashgate, 2006), pp. 205–23, who argues that there is a yet weaker, nontrivial version (that is, a nontrivial version that restricts the domain of properties quantified over even less).

in evaluating as candidates for the instantiation of thisness. I have already stated that I am primarily interested in whether events instantiate thisness, but since the standard counterexamples to (I/I) involve spatially dispersed material objects, it will be instructive for what follows if we first consider this class of individuals.

I.B.1. Spatial Dispersion. In the recent literature, the most commonly cited version of the argument from spatial dispersion is Max Black's famous (or infamous, depending on your point of view) thought experiment involving two iron spheres.[10] We are to imagine a world which has as sole occupants two, qualitatively indistinguishable, large, solid globes of iron. How do we know that there are two of them, if they share all of their qualitative properties? Owing, of course, to the fact that they are spatially dispersed. And, importantly, it is not only their nonrelational qualitative properties that they share, but also their relational ones, since all the relations that one bears to the other are born by the latter to the former: it is a perfectly symmetrical world. So, for example, each sphere has the property of "being two diameters from another iron sphere similar to itself."[11] Of course, there is a suppressed premise here, but it is one that would be difficult to resist; namely, that one and the same object could not possibly be in two different places at the same time. As others have done in this context, I will simply assume the truth of this premise.[12] So here we have a seemingly straightforward counterexample to (I/I): a possible world in which two material objects share all of their qualitative properties, yet are clearly distinct objects. And, given their distinctness, each globe must have some property in virtue of which it is the globe it is. That is to say, each globe must have primitive, nonqualitative thisness.

In section II, we will look at an important objection to this defense of the instantiation of thisness by material objects. For the moment, however, I want to look at two other objections which I believe are cause for less concern, but which merit a brief discussion nonetheless. First, one might worry that the leap from a possible world—in particular, a possible world far removed from the actual world—containing material objects with thisness, to the claim that actual material objects have thisness, is not warranted. I agree, however, with Adams, when he claims that, intuitively, the structure of individuality, whatever it might be, must be the same in all possible worlds.[13] In light of my approach here, of course, I would modify this claim by adding that it is the structure of individuality in particular classes of individuals that must be the same in all possible worlds. Thus, if in some possible world the structure of individuality for material objects is such that they are individuated by thisness, then, plausibly, that is the structure of individuality for material objects in all possible worlds.

---

[10] Black, "The Identity of Indiscernibles," *Mind*, LXI, 242 (April 1952): 153–64.

[11] Adams, "Primitive Thisness and Primitive Identity," p. 13.

[12] A notable exception is John O'Leary-Hawthorne, "The Bundle Theory of Substance and the Identity of Indiscernibles," *Analysis*, LV, 3 (July 1995): 191–96, where he argues that the bundle theory of substance can allow for identical, yet spatially dispersed, bundles of universals.

By my lights, the plausibility of this claim is grounded in the following line of thought. Individuation is determined by criteria of identity—criteria which vary according to the class of individual under consideration. Now, if these criteria of identity can serve to individuate a certain class of individuals in one possible world, then, given the necessity of the law of identity, they must do so in all possible worlds in which that class of individuals exist. So, the structure of individuality—for a particular class of individuals—must be the same in all possible worlds. For, suppose that this were not the case; then the spheres that are (allegedly) individuated by their thisnesses in a Black world, would not be so individuated were they to exist in a world *W* just like the actual world except that *W* also contains both of the spheres from the Black world (that is, the difference between the Black world and *W* is that *W* is not symmetrical, and thus the spheres in *W* do not share all their relational properties, as they do in the Black world). Given this, the spheres' qualitative properties would have to individuate them in *W*. But suppose that the spheres had precisely the same nonrelational qualitative properties in *W* that they had in the Black world. How could the very same properties that failed to individuate, that is, failed to identify, the spheres in the Black world, succeed in doing so in *W*? If the properties of 'being a sphere', 'having mass *m*', 'having radius *r*', 'being composed of iron', and so on, are unable to individuate the spheres in the Black world, then given the necessity of identity, neither would they be able to in *W*. The only other potential candidates for individuating properties are the different relational properties the spheres instantiate in *W* (in virtue of its asymmetry); but, *ex hypothesi*, these properties do not exist in the Black world, and so cannot individuate the *very same* spheres in *W*.[14]

Nevertheless, this defense of the argument from spatial dispersion leads to the second worry I mentioned—one that has been voiced by

---

[13] Adams, "Primitive Thisness and Primitive Identity," p. 13.

[14] There is, of course, another way to defend Adams's principle about the structure of individuality. By definition, thisnesses exist in all possible worlds in which the individuals that have them exist. So if thisnesses exist in the Black world, then they also exist (and, also by definition, individuate) in *W*. This defense, however, would not do much to persuade the opponent who is already skeptical of the concept of thisness.

John O'Leary-Hawthorne and J.A. Cover. They argue that the spatial dispersion argument suffers from an unwarranted bias in favor of taking transworld identity claims at face value.[15] Thus, in my example above, they would say that it is open to the opponent of nonqualitative thisness to deny that it is the very same spheres in the Black world and in *W*; for it might be the case that there are no determinate *de re* propositions true of either sphere in the Black world. Perhaps the only propositions true of the Black world are general ones such as $\exists x$ (x is a sphere), $\exists x \exists y$ (x is a sphere and y is a sphere and $x \neq y$) (*ibid.*, p. 104). Hence, though there are two spheres, there is nothing that there is for each sphere to be the particular sphere that it is—that is, they do not have thisnesses. And, given that there are no determinate *de re* propositions true of the spheres in the Black world, we cannot talk about one of those spheres existing in *W*. Transworld relations between qualitatively similar individuals are not structured in terms of the identity relation, but in terms of the counterpart relation. On this view, then, the possible existence of a Black world does not imply that thisness must be nonqualitative rather than qualitative; it implies that there is no thisness at all.

The lesson O'Leary-Hawthorne and Cover glean from these considerations is that an appropriately unbiased test for nonqualitative thisness must not only show that there is a pair of distinct individuals with all the same qualitative properties, but also must show that each member of the pair has different *de re* propositions true of them. I take it that strengthening the test in this manner is meant to dissuade actualists from defending their endorsement of transworld identity with an appeal to thisness, since it now appears that an assumption in favor of the former is required to establish the existence of the latter. For my part, I am happy to allow the more stringent test, as well as the assumption in favor of taking transworld identity claims at face value, since I am not trying to defend transworld identity. It would, of course, be unfortunate if any positive conclusions I come to about thisness—in particular, any conclusions that I might seek to employ in defense of a temporal framework—could only be accepted by those who endorsed transworld identity, but I am prepared to accept the inevitable exclusionary nature of systematic metaphysics. The systems generated from such an endeavor rarely produce a broad range of adherents, necessarily dependent as they are on a complex interweaving of doctrines, some of which will certainly be controversial. In any case, I take the assumption in favor of transworld identity to be considerably less controversial than some other assumptions one might make in arguing for thisness.[16]

---

[15] O'Leary-Hawthorne and Cover, "Framing the Thisness Issue," *The Australasian Journal of Philosophy*, LXXV, 1 (March 1997): 102–08.

I.B.2. Temporal Dispersion. Setting aside, then, those two objections, we will move on to the temporal dispersion counterexample to (I/I). Now that we have the general strategy in place for dispersion type arguments against (I/I), it will take less work to set up the one from temporal dispersion: we simply substitute events for material objects, and time for space. Before explaining how this works, however, I will need to say something about the conception of events I will be working with; since event theory is one of the more contentious areas of metaphysics, and my subsequent arguments will depend upon transparency with respect to my basic conception of what is to be an event. Nevertheless, I do not want to say anything now that would diminish the scope of those arguments, by, for example, assuming a specific (perhaps contentious) theory of events that turns out to be a necessary condition for accepting my ultimate conclusions. Thus, in order to steer a course between these two dangers, I will be explicit about what I minimally take an event to be, but I will refrain from offering or endorsing a fully worked out theory of events.

Minimally, then, I take an event to be a nonrepeatable, concrete particular. Of course, even this characterization of an event is not free from controversy, since some event theorists will claim that, inasmuch as distinct events can be co-located, they cannot be concrete. For my part, I take the concreteness of an event to consist in its being spatiotemporal and part of the causal nexus. Thus, I take this controversy to be grounded in the abstract/concrete distinction itself, rather than in different theories of events. The important point to take from my minimal characterization is that it rejects Chisholm's conception of events, according to which events are universal or repeatable entities.[17]

We now are in a position to discuss the argument from temporal dispersion. Again, the methodology is much the same as it is in the spatial dispersion case: if we can describe a possible world in which there are temporally dispersed events that share all of their qualitative properties, both relational and nonrelational, we will have provided a counterexample to (I/I) with respect to events. What would such a world look like? Hacking, quoting from A.J. Ayer, suggests a world consisting of "an infinite series of sounds ... A B C D A B C D A ..., succeeding one another at equal intervals, with no first or last term."[18] As I understand this example, we are to think of the terms 'A', 'B', 'C', and 'D' as each referring to an event with a particular set of qualitative properties. So, perhaps every occurrence of A has the property 'having pitch A'. For the sake of simplicity, and to avoid having to spell out a more thoroughgoing metaphysics of events, we will assume that the sounds and their properties are the sole constituents of these events. This assumption is in line with Ayer's original, minimal description of the example, as one involving just the sounds themselves, and not, for example, any material objects.[19]

---

[16] In section II, below, we will consider—and ultimately reject—two such controversial assumptions.

[17] See Roderick Chisholm, "Events and Propositions," *Noûs*, IV, 1 (February 1970): 15–24.

In elucidating and discussing the example, it will be helpful to make use of the type/token distinction. We will say that there are four event types in Ayer's world, where these types are individuated by the set of qualitative properties, both relational and nonrelational, that are instantiated by every token of that type. Thus, every occurrence of, for example, A, is a token of the event type 'A'; that is, an occurrence of an event with the properties 'having pitch A', 'being earlier than a token of event type 'B'', 'being later than a token of event type 'D'', and so on. It is immaterial, in this context, whether or not one wishes to think of the event types as universals, since the events themselves (that is, the token occurrences) are certainly not universals, and that is all my minimal conception of events requires. So, by stipulation, each of the token events in this world shares all of its nonrelational qualitative properties with all of the other tokens of its type; and the uniformity of the intervals, coupled with there being no first or last event, ensure that they also share all of their relational qualitative properties. Yet clearly, tokens of event type 'A', for example, are all distinct events, given that they are temporally dispersed. Thus, (I/I) fails for events in this possible world, and, given the preceding arguments, we may conclude that all events, both possible and actual, instantiate thisness.

II. AN OBJECTION TO DISPERSION ARGUMENTS

II.A. Hacking. I think the most powerful objection to the dispersion arguments we have just looked at comes from Hacking (*op. cit.*, pp. 249–56). His objection, and its application to these arguments, is important enough that we should take the time here to gain a thorough understanding of it; for it will have a significant bearing on the outcome of the investigation. Hacking objects not only to the dispersion arguments themselves, but to the whole strategy of inferring substantive metaphysical theses (such as the existence and instantiation of thisness) from supposed counterexamples to (I/I). His objection is that such examples can always be redescribed in a way that preserves all of the (nonquestion-begging) truths in the possible world under consideration, but in a way that does not imply that the qualitatively indistinguishable individuals in that world are distinct. It is important to note here that Hacking is not arguing for or against (I/I), he is simply arguing that the examples that are supposed to settle the matter are underdetermined.

---

[18] Hacking, "The Identity of Indiscernibles," this JOURNAL, LXXII, 9 (May 8, 1975): 249–56, p. 254.

[19] See Ayer, *Philosophical Essays* (New York: St. Martin's, 1954), p. 32.

Take the Black world, for example. According to Hacking, we are no less justified in claiming that there is only one sphere in that world, than we are in claiming that there are two.[20] All of the propositions, expressed in purely general terms, that are true of the world under the initial two-sphere description, are also true of the world under the one-sphere redescription—save the question-begging, "There are two spheres." According to Hacking, the latter proposition is question begging because asserting it is paramount to denying the truth of (I/I) without argumentation. Hacking anticipates the reply that "Every sphere is two diameters from a sphere" is neither question begging nor true in the one-sphere world, while it certainly is true in the two-sphere world. He acknowledges that this reply would be successful if space were absolute, since, on an absolutist view, relations such as 'being two diameters from a sphere' are not basic, but are determined by space itself. But if space is relational, and spatial relations are nothing more than relations between objects, then the totality of objects determines what spatial relations obtain. Thus, if space is relational, then "Every sphere is two diameters from a sphere" is true in the Black world in virtue of the truth of "There are two spheres"—and the latter is a proposition that we have agreed to omit from the Black world on the grounds that it is question begging.

So, according to Hacking, in order to demonstrate that there is a metaphysical difference between a Black world described as having one sphere and one described as having two spheres, we would first have to argue—on independent grounds—that space is absolute. One might think, of course, that absolute space is ruled out *ex hypothesi* in the Black world, since that world is set up as one in which nothing exists but the spheres (in which case there is no such entity as absolute space). This dismissal of absolute space might be question begging itself, however, so we will interpret Black's world as one in which no material objects exist but the spheres. Indeed, this must be the interpretation that Hacking makes, as he allows for the possibility of absolute space. Still, he claims that no good independent argument for absolute space is forthcoming, since it implies that a Black world where the spheres move a short distance along the "natural" axis is different from a Black world in which the spheres are stationary. Thus, "One should not commit oneself to such non-differences to show that a two-[sphere] universe differs from a one-[sphere] universe" (*op. cit.*, p. 252). Furthermore, once we allow that space is relational, then, as others have noted, we can redescribe the Black world as one in which space is curved, such that there is just one sphere that is spatially dispersed from itself; in which case, spatial dispersion does not imply distinctness.[21]

---

[20] I should note that Hacking actually discusses Kant's dispersion argument involving two raindrops abstracted from the actual world, rather than Black's argument from a two-sphere world.

As we have seen, Hacking also considers the temporal dispersion argument, and he makes essentially the same point as in the spatial dispersion case. We simply redescribe the world containing the series of events ... A B C D A B C D A ... as one that only contains A B C D (that is, for each event type, there exists only one token). This redescription is legitimate if time is relational, since temporal relations are then relations between events, and are determined not by the times at which they occur, but by the events themselves. Of course, inasmuch as this redescription implies that there is a first and last event, it is not faithful to the original description. Here, again, however, we import the notion of curved space-time, and claim that the world under the new description is circular: "with D preceding A, which precedes B, which precedes C, which precedes the very same D that precedes A."[22] Thus, though there are only four events under the redescription, the world is still faithful to the original description, since there is no first or last event.

II.B. Adams's Response. Adams, in responding to Hacking's objection, acknowledges that an acceptance of absolute space-time is too strong an assumption to make in defending dispersion arguments against (I/I).[23] He also acknowledges that, in the spatial case, the redescription of the Euclidean Black world as a non-Euclidean one in which space is curved does not make for a difference in possible worlds. If one were to claim that "a difference in geometries makes, in its own right, a difference in possible worlds," then one must assume that "what geometry the universe has does not depend on a determination of the number of objects in space" (*ibid.*, p. 16). Adams rejects this assumption on intuitive grounds, though he is not explicit about what those grounds are. I take it the thought is this: Just as, in our search for a nonquestion-begging counterexample to (I/I), we were unwilling to assume that relations between objects obtain independently of the number of objects in space (thus the ban on assuming absolute space), we should also deny the same assumption in the case of the universe's geometry. After all, for all we know, whether a world is Euclidean or non-Euclidean might have everything to do with how many objects it contains; and if it does, then we will have begged the question against (I/I) in assuming that the two-sphere world is different from the one-sphere world in virtue of being Euclidean. Thus, given that the above assumptions are questionable, and given that they are required in order to warrant the assertion of "There are two spheres" with respect to the Black world, Hacking is justified in claiming that this assertion is question begging.

---

[21] Though Hacking does not make the possibility of space-time curvature explicit in the spatial case, he does in the temporal case, as we shall see below.

[22] Hacking, *op. cit.*, p. 255.

[23] Adams, "Primitive Thisness and Primitive Identity."

In light of Adams's concessions to Hacking, how does he defend the argument from spatial dispersion? He does so by claiming that it is only plausible to regard the Euclidean and the non-Euclidean descriptions as indicating fundamentally different worlds if one has independent reason to believe that material objects have thisness; which he does not, given that he is "inclined to take a phenomenalistic view" of such objects (*ibid.*, p. 16n)! So much for our test for thisness—at least in the case of material objects—since *it* was supposed to give us independent reason for believing that such objects instantiated thisness.

Adams does, however, offer another brief argument in defense of the spatial dispersion counterexample. He considers the possibility of a world with *almost* indiscernible twin spheres. Here we are to imagine one sphere, for example, having a small chemical impurity that the other lacks. Such a world is clearly possible; and if a slight change in the chemical composition of the imperfect globe results in the two globes now being indiscernible, then, unless such changes affect the modal status of the world in question, the Black world should also be possible. I, however, find this argument inconclusive. Hacking's point was that the possibility of a two-sphere Black world hinges upon question-begging assumptions (this, in spite of the fact that such a world is conceivable), and that, absent those assumptions, Black worlds underdetermine the number of objects they contain. So, although the very same world with indiscernible twins can be redescribed as a single-sphere world, obviously the world with almost indiscernible twins cannot be so redescribed. In order for the two to be the same world, there must be something in the latter world to affect the change that brings about complete indiscernability. But it is hard to see what there could be in a world that, *ex hypothesi*, only contains iron sphere(s), that could possibly affect such a change. Perhaps, however, there need not be any entity that brings the required change about—perhaps it is just a brute change. Even allowing this, however—that is, even allowing that the chemical impurity of the one sphere is inexplicably removed at time *t*, rendering the two spheres indiscernible from *t*—the two spheres would still be discernible with respect to their distinct histories. Thus we are left with a world that is uncontroversially possible, but also one in which the spheres do not share all of their relational qualitative properties (they bear different temporal relations to their earlier states); in which case, it is no longer clear that thisness is required to distinguish them.

Perhaps, however, we should be considering two different possible worlds here, rather than a change in one. We have the world with the chemically impure sphere, which is clearly possible, and the world where the very same sphere is not chemically impure. So, given that the latter differs so slightly from the former, must not the latter also be possible? Again, this strikes me as being inconclusive. Clearly, if (I/I) turns out to be true in the case of material objects, then the world with indiscernible twins is not possible; in which case, small differences in chemical composition can make substantial modal differences. The point is, if one finds Hacking's objection compelling, then an appeal to the possibility of some world which is not a counterexample to (I/I), no matter how similar it is to the alleged counterexample, is not going to convince one of the possibility of the latter.

Adams does, however, go on to consider Black-like worlds inhabited with almost indiscernible twin human beings, and argues that the possibility of these worlds counts more conclusively against (I/I). And I grant that populating the spheres does make the argument from "almost indiscernibility" much more convincing, but it looks like material objects themselves, apart from any relations they might bear to human beings, have not conclusively passed the test for thisness. What about the temporal dispersion argument: Can it stand up any better against Hacking's objection? Or must our conclusions about thisness and events mirror those of thisness and material objects? It is this question to which we will turn next.

III. A WAY FORWARD: STATIC VERSUS DYNAMIC WORLDS

We have taken stock of the most serious objection to arguments purporting to establish the existence of thisness, and we have looked at responses to that objection which are, at best, inconclusive (at least when considered independently of the existence of human beings).

What, then, is the best way forward for the proponent of thisness? In particular, what is the best way forward for the proponent who suspects thisness might be a fundamental feature of reality—one that is independent of the existence of human beings? In the case of material objects, there appears to be no way forward. Given the reasonable ban on assuming absolute space, and given the equally reasonable ban on the assumption that a difference in geometry makes for a difference in possible worlds, there is no nonquestion-begging way to infer the instantiation of thisness in material objects from spatial dispersion arguments. In the case of events, however, there is a way forward, and the key to recognizing it lies in recognizing an important disanalogy between space and time.

Recall the intuitive reasoning behind rejecting the assumption that a difference in geometry makes for a difference in possible worlds: whether or not a Black world is Euclidean or non-Euclidean may very well depend on whether or not it contains two spheres or only one. Thus we cannot validly infer that a Black world contains two spheres simply from the stipulation that it is Euclidean. What about the temporal case? Are we equally obligated to accept a ban on the assumption that a difference in chronometry makes for a difference in possible worlds? Given the ban on assuming absolute space-time, we cannot claim that the number of events in a temporal series are determined by relations to the times at which they occur; but must we also reject the thought that the number of events in a temporal series might be determined by whether that series is linear or circular? If we take a moment here to reflect on the nature of circular time, I think it will reveal that a difference in chronometry does make for a difference in possible worlds.

Consider some of the necessary characteristics of circular time: First, as we have already noted, there is no first or last event in circular time. It might be contended that this is not a necessary characteristic, but is instead a contingent characteristic stipulated in order to remain faithful to the original description of Ayer's linear repeating time. But if we think of the series of events in circular time as forming "a seamless, closed circle," as Richard Sorabji urges in his discussion of circular time, then it is clear that there is no nonarbitrary way of designating a first or last event.[24]

Second, although there is a sense in which the events in a circular time series continually repeat themselves, strictly speaking, each event occurs only once—given that each event is numerically identical to all of its "repetitions." Again, the type/token distinction can help clarify the distinction between Ayer's linear repeating world and the corresponding circular world: both worlds contain only four event types, but Ayer's world also contains an infinite number of tokens of those types; whereas the circular world only contains four tokens—one corresponding to each type. It is also helpful here to think of events in circular time on analogy with the numbers of a clock face. Just as 12 o'clock only occurs once on a clock face, so each event in circular time only occurs once (*ibid.*).

---

[24] Sorabji, *Matter, Space, and Motion* (London: Duckworth, 1988), p. 165.

Third, given the first and second characteristics, all events in a circular time series must be equally real. This follows because if each of the four events in the circular time series A B C D occurs once and only once, and there is no beginning to the series, then it is always the case that every event has occurred (which, of course, is just to say that every event *occurs* (tenselessly)). And I know of no more apt criterion for the reality of an event than its occurrence. So when event A is present to a subject, it is also the case that that very same event lies in the subject's past and future, and thus there can be no objective distinction between past, present, and future in circular time. And finally, it follows from all this that a circular time series is necessarily static. By this, I mean that a circular temporal world is necessarily one in which the B-Theory of time obtains.[25]

Now consider linear repeating time. Clearly, it is conceivable that such a time series be static, but it should be just as clear that it is also conceivably (that is, conceptually possibly) dynamic. Certainly, *actual* linear time is conceivably dynamic—by which I mean that holding actual linear time to be dynamic does not involve any logical or conceptual incoherence (*pace* John McTaggart[26]). This is precisely why there exists such a lively debate between the static and dynamic theorists: there is no way logically or conceptually to demonstrate the falsity of either theory. In contrast, it is my contention that if time were circular, one *could* conceptually demonstrate the falsity of the dynamic theory; since it is conceptually incoherent to claim that (i) all events exist tenselessly (as I argued, above, that circular time entails), and (ii) tense is an objective feature of reality (as the A-Theory entails). Thus, if actual linear time is conceivably dynamic, then I can see no reason to deny that the repeating linear time of Ayer's world is also conceivably dynamic. After all, the differences between Ayer's world and the actual linear world are not ones that obviously entail the incoherence of the dynamic theory in the former.

---

[25] For my purposes here, I take the B-Theory of time to denote a static conception of time according to which there is no objective temporal becoming, and according to which past, present, and future events all tenselessly exist. I take the A-Theory to denote a dynamic conception of time according to which temporal becoming (however conceived) and tense are objective features of reality.

[26] Even if one were to acknowledge the alleged contradiction of the A-series that McTaggart argued for, there is nothing intrinsic to linear time that entails the contradiction. Consider that all sides to the debate agree that presentism is a dynamic theory that is immune to the contradiction, and presentism normally assumes linear time.

As an example of what an A-Theoretic Ayerian world might look like, suppose the linear world that contains the series ... A B C D A B C D A ... is a dynamic one, according to which temporal becoming is an objective feature of reality. Further suppose that on this particular dynamic conception of time, each event causes its successor to come into existence. So, for example, when A begins to occur, the subsequent occurrence of B does not yet exist.[27] Now compare this with static circular time, where A, B, C, and D always exist. We have here a very real difference in possible worlds, given that the two have different ontologies; and, importantly, this difference in ontology is not determined by questionable assumptions about absolute space-time, nor by question-begging assumptions that *presuppose* a difference in ontology. On the contrary, the difference in ontology follows *from* the nonquestion-begging descriptions of a world as linear and a world as circular. Given these descriptions, which both parties to the debate acknowledge, our goal has been to determine whether they indicate two different worlds, or whether the linear world can be equally well described as a circular world. What we have discovered is that a circular world is necessarily static, while a repeating linear world is not, and that this difference determines a difference in ontology. So although a difference in geometry does not make for a difference in possible worlds, a difference in chronometry does. And the disanalogy between space and time that underpins this conclusion is that, while we can conceive of time being static or dynamic, we can only conceive of space as being static; thus there is no corresponding ontological distinction to be made when comparing Euclidean and non-Euclidean spatial worlds.[28]

---

[27] As will become evident, the line of argument I am pursuing here could be simplified by stipulating a "moving spotlight" dynamic conception for the linear world, rather than a nonexistent future dynamic conception. The former acknowledges the reality of past and future events, but weds this ontology with an objective temporal becoming, that is, the moving spotlight of the present. The benefit of simplicity in this approach, however, is outweighed by a greater cost: the moving spotlight conception fails to provide a robust enough distinction between past and future; and it is precisely such a distinction that best supports the claim that there is a metaphysical difference between the circular and linear worlds. For a defense of the view that a dynamic conception of time requires an existent past and nonexistent future, see my "Presentism and Ontological Symmetry," *The Australasian Journal of Philosophy*, LXXXIII, 2 (June 2005): 223–40.

[28] I have put the modal claims in the foregoing discussion in terms of conceptual—rather than metaphysical—possibility and necessity. I take conceptual necessity to be narrower than metaphysical necessity, and strictly logical necessity to be narrower

Let us now consider possible objections to this line of argumentation. First, one might object to the proffered dynamic conception of infinitely repeating time on the grounds that a nonexistent future cannot be infinite. But we can certainly conceive of this series continuing to infinity without it being the case that—at the time of an event's occurrence—any of its successors exist. Whether or not there could be an end to time is an entirely independent question from whether time is static or dynamic. It is not as though a philosopher who endorses a dynamic conception of time and a nonexistent future is thereby forced to deny that time continues to infinity.

Second, one might argue that a nonexistent future implies that the different repetitions of each event have different qualitative properties; in which case, we do not get a counterexample to (I/I).[29] The worry would be that a given occurrence of A, for example, differs from previous occurrences perhaps because it has different tensed properties, or perhaps because it bears no future directed relations to other events; and if either or both of these differences are qualitative ones, then (I/I) succeeds in differentiating between distinct occurrences of A, thus nullifying the justification for postulating thisness. My response to this objection is, first, to note that if the relations that are meant qualitatively to distinguish a present event from its previous occurrences are relations between particular events, then they are not qualitative relations at all and we still have a counterexample to (I/I). Furthermore, the same point can be made in connection with times: relations to particular times cannot qualitatively distinguish between event repetitions. If, however, the properties and relations that are meant qualitatively to distinguish a present event from its previous occurrences are characterized in purely general terms, then I acknowledge that for my argument to succeed, these properties and relations must fail to distinguish between event repetitions.

First, let us consider whether tensed properties can distinguish between event repetitions. Take, for example, the following propositions,

---

still. These grades of necessity assume a tradition of modality that can be found in Alvin Plantinga, *The Nature of Necessity* (New York: Oxford, 1974), p. 2, and Jonathan Lowe, *The Possibility of Metaphysics* (New York: Oxford, 1998), pp. 13–16. It is, in my view, an advantage of this tradition that it allows for the conceptual possibility of both the A-Theory and B-Theory (and so meaningful debate), while also allowing that the theory that actually obtains is metaphysically necessary (that is, obtains in all logically possible worlds), and that the one that does not actually obtain is metaphysically impossible.

[29] Let "repetitions" in this context mean "different token occurrences of the same event type."

each of which ascribes a tensed property to every event in our dynamic, linear repeating time series:

(1) Every event is past at some time.
(2) Every event is present at some time.
(3) No event is future at any time.

Since (1), (2), and (3) are all true in the world under consideration, their constituent tensed properties fail to differentiate between event repetitions.

What about temporal relations between events in the world under consideration: Do they qualitatively distinguish between event repetitions? At this point in the debate, one might just appeal to the nature of a dynamic world and claim that such relations are not basic, but are reducible to tensed properties. This response, however, fails to take into account the possibility that a dynamic world with a lopsided ontology, such as the one currently under consideration, requires that tenseless temporal relations be basic and irreducible to tensed properties.[30] Given this possibility, I wish to remain agnostic on whether tensed properties or tenseless temporal relations are more basic. So let us consider

(4) Every event is earlier than another event at some time.

Once again, (4) is a purely general proposition with no particular constituents, and, once again, given that it is true in the world under consideration, it fails to differentiate between events. Now, however, consider

(5) Every present event is earlier than another event.

(5) is false in the stipulated world, and so, if it is purely general, its constituent relation does qualitatively distinguish different event repetitions. But is (5) purely general? On any dynamic view, given the objectivity of temporal becoming, what is present is objective and unique, and so there is only ever one present event. Therefore, 'present event' denotes a particular event, and thus (5) is not purely general. A similar point can be made against

(6) Every event is earlier than another event at the present time.

Although (6) is false in the world under consideration, it is not purely general, given that, on a dynamic view, the 'present time' is a particular time. Likewise for any proposition that has the 'the present' as a constituent, such as

---

[30] See Michael Tooley, *Time, Tense, and Causation* (New York: Oxford, 1997). On Tooley's view, though tensed properties and concepts are not basic, the concept of actuality as of a time is basic, thus preserving a dynamic world.

(7) Every present event is three event occurrences earlier than a similar event.

Again, this proposition, though false in the dynamic, linear repeating world, makes reference to a particular event (the 'present event'), in which case the world under consideration remains a counterexample to (I/I).[31] In a dynamic world, it is an objective fact about which particular event is present at which particular time, so propositions in such a world cannot contain references to 'the present' and remain general. It is important here to be clear about what this claim entails, and what it does not entail. It does not entail that there are not different ways of carving up the temporal boundaries of the present. What it does entail is that whatever way we choose to stipulate those boundaries, and so designate a present event, that way must make some reference to the edge of becoming. On the view under consideration, the edge of becoming is the time *t* at which events come into existence: prior to *t* the designated event does not exist, and subsequent to *t* it does exist. It is the reference to this edge of becoming that makes propositions about the present singular rather than general.

There are no doubt many other potential candidates for propositions that are not true of all events in the dynamic, linear repeating world, but I maintain that any such propositions would fail to be purely general, and thus would be unable to qualitatively distinguish between different tokens of the same event type. I therefore conclude that this linear world does provide a counterexample to (I/I), and that the objection under consideration therefore fails. In light of that failure, we can stand by the claim that a circular world is necessarily static, while a repeating linear world is not, as well as the conclusion that this constitutes a real difference in possible worlds.

IV. CONCLUSION

We have been trying to discover whether there are thisnesses in the world, and, in particular, whether events instantiate them. Along the way we also looked at arguments for the view that material objects instantiate thisness. What we have learned is that, unless we are willing to endorse absolute space-time, we are unable to successfully argue that material objects, in and of themselves, instantiate thisness. Initially, it also appeared that the same considerations which cast doubt on thisness for material objects weigh against thisness for events as well. Then, however, we noticed an important disanalogy between space and time: a temporal series might be static or dynamic, but space can only be static. This disanalogy gave us room to assert that temporal dispersion cases are definitive, nonquestion-begging counterexamples to (I/I), even though spatial dispersion cases are not. This is because it is no arbitrary matter how many events exist in the temporal dispersion cases. If one such case is described as a dynamic, linear repeating temporal world with infinitely many events, then one cannot redescribe that very same temporal world as a circular one with only four events, since circular time is necessarily static, and a static temporal world is an altogether different one from a dynamic temporal world. Therefore, given that temporal dispersion cases—even ones not involving human beings—are counterexamples to (I/I), events do have thisness, and their instantiation of thisness is logically and metaphysically independent of the existence of human beings.

---

[31] I hope it is clear that this proposition is also open to the same objection that Hacking alleged, in the spatial case, against, 'Every sphere is two diameters from another sphere'. That is, it assumes that we can determine the number of events in a temporal series from the relations that hold between times, and this assumes absolute time. Cleary this is not an assumption that the opponent of thisness would be willing to make, since, as we have seen, the spatial and temporal dispersion arguments provide clear cut counterexamples to (I/I) if absolute space-time is assumed.

If this last conclusion is correct, and the temporal dispersion argument goes through, then am I not being a little pessimistic about the spatial dispersion argument? Perhaps I am simply overstating the disanalogy between space and time, since one might wish to question whether our inability to conceive of dynamic space is really indicative of the actual, underlying ontology, or whether that inability is merely a limitation of our conceptual framework. These questions might encourage us to run the argument in the opposite direction. Instead of diagnosing our success in the temporal dispersion argument, and failure in the spatial dispersion argument, as involving a disanalogy between space and time, we could cite the close *analogy* between space and time, and our success in the temporal dispersion argument, as warrant for believing there must be a successful version of the spatial dispersion argument.

But what would such an argument look like? It would either have to propose a coherent conception of what dynamic space would involve, or it would have to propose some other, particularly spatial—perhaps geometrical—distinction, that was analogous to the particularly temporal distinction of static/dynamic. Furthermore, bear in mind that such a proposal on its own would not suffice to rescue the spatial dispersion argument, since the successful proposal would also have to entail, in a nonquestion-begging way, that there are two spheres, rather than one, in the Black world. As we saw in section II, assuming absolute

CAUSAL INDEPENDENCE, THE IDENTITY OF INDISCERNIBLES, AND THE ESSENTIALITY OF ORIGINS*

Max Black offers the following counterexample to the Principle of the Identity of Indiscernibles (PII):

(Passage I)

> Isn't it logically possible that the universe should have contained nothing but two exactly similar spheres? We might suppose that each was made of chemically pure iron, had a diameter of one mile, that they had the same temperature, colour, and so on, and that nothing else existed. Then every quality and relational characteristic of the one would also be a property of the other. Now if what I am describing is logically possible, it is not impossible for two things to have all their properties in common. This seems to me to refute the Principle.[1]

Commenting on Black's example, Robert Adams uses the possibility of almost indiscernible spheres to argue for the possibility of Black's indiscernible spheres:

(Passage II)

> We may just have an intuition that there could be distinct, though indiscernible, globes in these circumstances. But there may also be an argument for this view—which will depend in turn on other intuitions, like all arguments in these matters. The argument might rest on an intuition that the possibility of there being two objects in a given spatiotemporal relation to each other is not affected by any slight changes in such features as the color or chemical composition of one or both objects. If we accept that intuition, we can infer the possibility of indiscernible twins from the uncontroversial possibility of *almost* indiscernible twins. No one doubts that there could be a universe like the universe of our example in other respects, if one of the two globes had a small chemical impurity that the other lacked. Surely, we may think, the absence of the impurity would not make such a universe impossible.[2]

Even those who are not convinced by Adams's argument, such as Michael Della Rocca,[3] acknowledge the extent of its-influence, yet

---

*I thank James Sennett, Yuri Balashov, and Randolph Clarke for their feedback on an earlier version of this paper.

[1] Black, "The Identity of Indiscernibles," *Mind*, LXI (1952): 153–64, at p. 156.

[2] Adams, "Primitive Thisness and Primitive Identity," this JOURNAL, LXXVI, 1 (January 1979): 5–26, at p. 17.

[3] See Della Rocca, "Two Spheres, Twenty Spheres, and the Identity of Indiscernibles," *Pacific Philosophical Quarterly*, LXXXVI (2005): 480–92.

0022-362X/09/0605/277–91

the argument itself is not fully spelled out in the passage. Why is it that "Surely...the absence of the impurity would not make such a universe impossible"? The same question arises with respect to the second "almost indiscernibles" argument that Adams discusses. Adams postulates a world in which he and an almost indiscernible twin differ only in how many horns appear on the heads of the dragons in their corresponding nightmares. The dragon in Adams's nightmare has ten horns, whereas the dragon in the twin's nightmare has only seven. Adams writes as follows:

> But if such a world is even possible, it seems to follow that a world with perfectly indiscernible twins is also possible. For surely I could have existed, and so could my twin, if my monster had had only seven horns, like his (*op. cit.*, pp. 17–18).

Why is it that "...surely I could have existed, and so could my twin, if my monster had had only seven horns, like his"? Adams acknowledges (*op. cit.*, p. 18) that his argument depends on the necessity of identity and distinctness and on the postulation of transworld identities, among other things, but the additional assumptions Adams identifies do not get us from the premise to the conclusion. Adams's 'surely's remain. My aim in this essay is to get rid of them.

I will begin by formulating the conclusion and an explicit premise of Adams's argument in Passage II. I will then show that the gap in this argument can be bridged using entirely plausible assumptions about causal independence and the logic of the counterfactual conditional. I will then argue that causal independence bears in a similar way on an issue Nathan Salmon raises[4] concerning Saul Kripke's argument[5] for the Essentiality of Origins. I will conclude by using my analysis of the argument of Passage II to identify a previously unacknowledged reply to that argument.

I. THE ARGUMENT OF PASSAGE II

Let Sx abbreviate the claim that x is an iron sphere one mile in diameter. We can suppose that Sx is a formula describing x as having a specific internal structure, for example, a specific number of iron atoms arranged in a specific way. Let Mxy be the claim that (i) x and y exist,[6] and (ii) whatever exists is either x, y, a part of x, a part of y, or something the parts of which come only from x and y, and (iii) the surface of x is two miles from the surface of y. Finally, let Ix

---

[4] See Salmon, *Reference and Essence* (Princeton: University Press, 1981).

[5] See Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980).

[6] To be more specific, (i) is the claim ($\exists z\ z=x\ \&\ \exists z\ z=y$).

abbreviate the claim that *x* has a chemical impurity bonded to its surface. The inference in Passage II from the possibility of the almost indiscernible spheres to the possibility of the indiscernible spheres seems to be the inference from this premise:

(1) $\Diamond\, \exists x\, \exists y\, [(Sx\ \&\ Sy)\ \&\ (Ix\ \&\ {\sim}Iy)\ \&\ Mxy]$

to this conclusion:

(2) $\Diamond\, \exists x\, \exists y\, [(Sx\ \&\ Sy)\ \&\ ({\sim}Ix\ \&\ {\sim}Iy)\ \&\ (Mxy\ \&\ x{\neq}y)]$

In neither the premise (1) nor the conclusion (2) does the formula inside the quantifiers completely describe the spheres *x* and *y* or explicitly state their indiscernibility. The conclusion does, however, state that *x* and *y* are numerically distinct while having the properties postulated by Black in Passage I. The inference from (1) to (2) is invalid in the relevant logical system, namely quantified S5,[7] since (1) is consistent with (3) whereas (2) is not:

(3) $\Box\, \forall x\, \forall y\, [((Sx\ \&\ Sy)\ \&\ ({\sim}Ix\ \&\ {\sim}Iy)) \supset x{=}y]$

That is, (1), but not (2), is consistent with its being impossible for more than one chemically pure iron sphere to exist. So, if the argument of Passage II is valid, it must have tacit premises. What are they?

Adams's comments in the following passage are relevant to the search for tacit premises. Switching from the iron spheres to the dreaming twins, suppose that *w* is the world in which Adams and his twin are almost indiscernible, that is, indiscernible except for the numbers of horns on the monsters in their respective nightmares:

> ([A]) The nonidentity obtaining between me and my twin in *w* is proved by a qualitative difference between us there. ([B]) The argument depends on an intuition of transworld identity that in a possible world (call it *w′*), otherwise like *w*, but in which my dragon has only seven horns, there could exist an individual identical with me and an individual identical with my twin, even though we would not be qualitatively different in that case. ([C]) The transitivity of identity is relied on in arguing that since my twin and I are not identical in *w* (as shown by the difference in our suchnesses there), it follows that we are not identical in any possible world, and therefore are distinct in *w′*, if we both exist in it (*op. cit.*, p. 18).

---

[7] By *quantified S5* I mean a logical system that combines propositional S5 with the quantifiers in a way that validates the Necessity of Identity but not the Barcan Formula or its Converse.

Adams formulates these comments in terms of his twins example, but the same points apply to the spheres example. Switching from the twins to the spheres, Adams's (B) suggests that (4) should be regarded as an intermediate step in the argument for (2):

(4) $\Diamond\, \exists x\, \exists y\, [(Sx\ \&\ Sy)\ \&\ (Ix\ \&\ {\sim}Iy)\ \&\ Mxy\ \&\ \Diamond[(Sx\ \&\ Sy)\ \&\ ({\sim}Ix\ \&\ {\sim}Iy)\ \&\ Mxy]]$

That is, it is possible for there to be iron spheres *x* and *y* existing alone and two miles apart such that while *x* but not *y* has a chemical impurity, it is possible for these same iron spheres *x* and *y* to exist alone two miles apart both free of any impurity. The transworld identification that Adams refers to in (B) is reflected in (4) by the use of *x*- and *y*-quantifiers to quantify into the scope of the second possibility operator ◇. As regards the idea that (4) is an intermediate step in the argument from (1) to (2), the good news is that (2) does follow from (4) in quantified S5. The bad news is that since (2) follows from (4) whereas (2) does not follow from (1), (4) does not follow from (1), and Adams's discussion does not explain how the gap between (1) and (4) should be bridged any more than it explains how the gap between (1) and (2) should be bridged. What Adams's endorsement of transworld identity in (B) does suggest is that the path from (1) to (2) goes through (4) in the sense that (1) and whatever tacit premises we identify should entail not only (2) but (4), as well.

II. BRIDGING THE SURELY-GAP

Consider the almost indiscernible spheres that Adams postulates in Passage II. These spheres may exist in an atmospheric vacuum, but that does not mean that they exist in a *nomological* vacuum. The world of Passage II can be assumed to have some laws of nature or other, and this fact can be exploited in our attempt to bridge the gap in the argument of Passage II.

What laws of nature should the world of the almost indiscernible spheres be assumed to have? At first blush, nothing in Passage II seems to indicate that this world is nomologically exotic. Can it then be stipulated to have the very same laws of nature as the actual world? Let us see what such a stipulation would entail.

One of the features of the laws of nature in the actual world is that they permit a great deal of *causal independence*. In particular, in the actual world, it is easy to find or arrange a situation in which certain of the characteristics of one piece of matter are causally independent of certain of the characteristics of other nearby pieces of matter. Consider my neighbor and me and our cars. My car is dent-free, but my neighbor was in a minor accident, and his car was dented.

My car was nowhere near the scene of the accident at the time, so it was not a case of a third driver swerving to avoid me and hitting my neighbor instead. Clearly, my car's being dent-free is causally independent of whether my neighbor's car was dented. Now, what if my neighbor's car had not been dented? The laws of nature could have been such that, for example, a near miss on my neighbor's car would produce a dent in my car through a law of action-at-a-distance, but, in the actual world, that is not how things work. The dent-free state of my car is causally independent of whether my neighbor's car was dented, and this independence (which is a matter both of the laws of nature and the contingent facts about our cars) manifests itself in the truth of the following counterfactual conditional: *My car would still have been dent-free if my neighbor's car had not been dented.*

One might argue, by parity of reasoning, that if Adams's world containing the two almost indiscernible spheres shares the actual world's laws of nature, then Adams's world is one in which the chemically pure state of the chemically pure sphere is causally independent of whether the other sphere has a chemical impurity. If so, then this causal independence manifests itself in the truth of the following counterfactual conditional: *The sphere that has no impurity would remain chemically pure if the other sphere also had no impurity.* This, if true, seems to give us reason to conclude that the two spheres would be indiscernible if the chemically impure sphere instead had no impurity. Since the antecedent of this counterfactual is possibly true, so is the consequent, and the possibility of Black's world is secured. Or so the argument goes.

It is not clear that the line of reasoning just described will succeed. In particular, the appeal to parity of reasoning may be too much of a stretch. Adams's almost-indiscernibles world is quite a different place from the streets where my neighbor and I drive our cars, and it was no part of the two-cars example that our cars would be indiscernible if my neighbor's car had not been dented. And how did the almost indiscernible spheres in Adams's world come into existence if Adams's world is governed by the actual world's laws of nature? If the two spheres are just two miles apart and obey the actual world's laws of nature, will gravity not soon pull them into a collision (perhaps giving them distinguishing dents of their own)? And can we be sure how alike the two spheres would be if the blemished one had no impurity? In the end, it is an empirical question whether the actual world has laws of nature that support the causal independence of any given pair of propositions about the two spheres. No matter; the world of the almost indiscernible spheres need not be stipulated to have precisely the same laws of nature as the actual world. It will be enough, for the purposes of the argument of Passage II, if the world of the almost indiscernible spheres has laws of nature that support the relevant claim of causal independence. Indeed, that world need have only one law of nature, namely, that whenever two almost indiscernible spheres are as described in Passage II, the relevant causal independence obtains. Thus, the laws of nature in the world of the almost indiscernible spheres may be somewhat exotic after all, but, exotic or not, it is completely plausible that a world with such laws is metaphysically possible.[8]

The upshot? The world of the almost indiscernible spheres has been under-described. To describe it properly and explain in detail (as we will below) how its possibility entails the possibility of Black's world, we need additional conceptual resources, namely the counterfactual conditional and the related notions of laws of nature and causal independence. Once the facts about laws of nature and causal independence in Adams's world are taken into account, the argument for the possibility of Black's world can be seen to be valid—and no longer in need of the 'surely's.

II.1. Causal Independence and Counterfactuals. Gibbard and Harper offer the following definition of the causal independence of a state on the choice between two actions:

> To say that a state S_i is causally independent of the choice between A and B is to say that S_i would hold if A were performed iff S_i would hold if B were performed.[9]

This definition concerns the independence of a state from a choice, but there is no relevant difference between states, choices, and propositions that prevents the definition from being generalized to a definition of causal independence for a pair of propositions p and q. The result would be the following counterfactual analysis of the claim $N(p,q)$ that the truth of p is causally independent of the truth or falsity of q, where '>' is the counterfactual or subjunctive conditional:

$$(5)\ N(p,q) =_{\mathrm{df}} (q > p) \equiv (\sim q > p)$$

That is, $N(p,q)$ means that p would be true if q were true, if and only if p would be true if $\sim q$ were true. $N(p,q)$ captures the idea that a supposition about the truth value of q would make no difference to whether p would be true: either p would be true either way, or else p might[10] be false either way.

---

[8] I thank Yuri Balashov and James Sennett for helping me sort out the argument of this paragraph.

[9] Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility," in William L. Harper, Robert Stalnaker, and Glenn Pearce, eds., *Ifs* (Boston: Reidel, 1978), pp. 153–90, at p. 172.

To get an appreciation of what (5) entails, let us next introduce some standard assumptions about the logic of the counterfactual conditional and its interaction with the logic of metaphysical necessity and possibility:[11]

(6) $\vdash (p \,\&\, q) \supset (p > q)$

(7) $\vdash (p > q) \supset (p \supset q)$

(8) $\vdash \Diamond p \supset ((p > q) \supset {\sim}(p > {\sim}q))$

(9) $\vdash \Box(p \supset q) \supset (p > q)$

(10) $\vdash \Diamond p \equiv {\sim}\Box{\sim}p$

(11) If $p_1, \ldots, p_n \vdash q$, then $(p > p_1), \ldots, (p > p_n) \vdash (p > q)$.

(12) If $p_1, \ldots, p_n \vdash q$, then $\Box p_1, \ldots, \Box p_n \vdash \Box q$.

(13) If $\vdash q$, then $\vdash \Box q$.

(14) If q is a tautology, then $\vdash q$.

(15) If q is a tautological consequence of $p_1, \ldots, p_n$, then $p_1, \ldots, p_n \vdash q$.

(16) $\vdash$ is compact, reflexive, monotone, and transitive.

Conditions (6)–(16) hold, for example, both for David Lewis's preferred counterfactual logic[12] and for Robert Stalnaker's preferred conditional logic.[13] Given (5) and (6)–(16), we have the following:

(17) $N(p,q), p, q \vdash {\sim}q > p$

(18) $N(p,q), p, {\sim}q \vdash q > p$

(19) $N(p,q), p, q, \Diamond{\sim}q \vdash {\sim}({\sim}q > {\sim}(p \,\&\, {\sim}q))$

(20) $N(p,q), p, {\sim}q, \Diamond q \vdash {\sim}(q > {\sim}(p \,\&\, q))$

---

[10] I adopt Lewis's 'might'-conditional idiom for the purposes of this discussion; see David Lewis, *Counterfactuals* (Cambridge: Harvard, 1973). That is, I am translating ${\sim}(p > {\sim}q)$ as "Were it the case that p, it might be the case that q." A more neutral translation of ${\sim}(p > {\sim}q)$ would be "Its being the case that p would not rule out its being the case that q."

[11] Where X and Y are sets of statements, to say that $\vdash$ is compact means that if $X \vdash p$, then, for some finite $Y \subseteq X$, $Y \vdash p$. To say that $\vdash$ is reflexive means that $X \vdash p$ if $p \in X$. To say that $\vdash$ is monotone means that if $X \vdash p$ and $X \subseteq Y$, then $Y \vdash p$; to say that $\vdash$ is transitive means that if $X \vdash p$ and $X \cup \{p\} \vdash q$, then $X \vdash q$.

[12] See Lewis, *op. cit.*

[13] See Stalnaker, "A Theory of Conditionals," in Nicholas Rescher, ed., *Studies in Logical Theory*, American Philosophical Quarterly, Monograph 2 (Oxford: Blackwell, 1968), pp. 98–112.

CENTRAL LIBRARY UNIVERSITY OF CALCUTTA

Definition (5) is actually a stronger assumption than is needed for the points to be made below. It would suffice to forego (5) as a definition of causal independence, to instead take causal independence as primitive, and to assume (19) and (20) as principles relating causal independence to metaphysical possibility and the counterfactual conditional. This would make my argument consistent with the view that causal independence cannot be given a counterfactual *analysis*.

What (19) says is that given that p is causally independent of whether q, and given that p and q are true and $\sim q$ is possibly true, it follows that *the truth of* $\sim q$ *would not rule out the joint truth of* p *and* $\sim q$, or, to use Lewis's idiom, p and $\sim q$ *might both be true if* $\sim q$ *were true.* What (20) says is that given that p is causally independent of whether q, and given that p and $\sim q$ are true and q is possibly true, it follows that *the truth of* q *would not rule out the joint truth of* p *and* q, or, to use Lewis's idiom, p and q *might both be true if* q *were true.* By (19), (20), and (6)–(16), we have the following:

(21) $N(p,q), p, q, \Diamond\sim q \vdash \Diamond(p \;\&\; \sim q)$

(22) $N(p,q), p, \sim q, \Diamond q \vdash \Diamond(p \;\&\; q)$

What (21) says is that given that p is causally independent of whether q, and given that p and q are true and $\sim q$ is possibly true, it follows that the joint truth of p and $\sim q$ is also possible. That is, as long as $\sim q$ is at least possible, and p and q are both true, the causal independence of p from whether q ensures that p and $\sim q$ are compossible. (22) makes a similar point: as long as q is at least possible, and given the truth of p and $\sim q$, the independence of p from whether q ensures that p and q are compossible.

II.2. Almost Indiscernible Spheres with Causal Independence. Returning to the argument of Passage II, our point earlier about the laws of nature in Adams's world comes to this: the almost indiscernible spheres can be assumed to exist in a world in which certain laws of nature (abbreviated as L) hold, where L entails that if x and y are any iron spheres of a certain structure situated alone and two miles apart, where x has a chemical impurity though y does not, then the fact that *iron sphere* y *is chemically pure and located two miles from iron sphere* x *in circumstances in which nothing else exists* is causally independent of whether x *is chemically impure*:

(23) $\Box(L \supset \forall x \forall y\, [(Sx \;\&\; Sy \;\&\; Ix \;\&\; \sim Iy \;\&\; Mxy) \supset N((Sx \;\&\; Sy \;\&\; \sim Iy \;\&\; Mxy), Ix)])$

Factoring L into (1) and adding that in the almost-indiscernibles world it is possible for the impure sphere x to be free of any impurity, we obtain the following, which now replaces (1) in a revised analysis of the argument of Passage II:

$$(24)\ \Diamond\ \exists x\ \exists y\ [(Sx\ \&\ Sy)\ \&\ (Ix\ \&\ {\sim}Iy)\ \&\ Mxy\ \&\ (L\ \&\ \Diamond{\sim}Ix)]$$

Finally, where $\vdash$ is logical consequence in quantified S5 extended to include a conditional logic that satisfies (21) and (22), both the "intermediate" conclusion (4) and the final conclusion (2) are derivable from the tacit premise (23) and the revised explicit premise (24).[14] In other words, the possibility of Black's indiscernible spheres world follows from the possibility of Adams's almost-indiscernible-spheres world once we recognize the cooperating roles of causal independence and laws of nature in mediating the inference. No more 'surely's.

II.3. It Is Not a Continuity Argument. Some commentators, such as Peter Forrest, have suggested that the argument of Passage II is a Continuity Argument:

> A rejoinder to [a certain defense of PII] is the continuity argument, essentially due to Adams (1979). It is granted that almost perfect symmetry is possible. For there could be a space with nothing in it but two distinct spheres differing very slightly. Black's example of qualitatively identical spheres is the limiting case as the differences get less and less.[15]

Gonzalo Rodriguez-Pereyra gives a similar analysis:

> But Black's world is possible. For there is a possible world with two almost indiscernible spheres. That is, there is a possible world like Black's except that the spheres differ infinitesimally in temperature.... That world contains two particular spheres, *a* and *b*. But if *a* has a temperature *T* and a different particular *b* of the same kind as *a* has a temperature T^* infinitesimally different from *T*, then it is possible for *a* to have T^*. Thus if the world with the almost indiscernible spheres is possible, so is another world in which the spheres are completely indiscernible.[16]

The possibility of the indiscernibles-world is justified, on this view, by the fact that for any positive degree of qualitative difference, no matter how small, there is an almost-indiscernibles-world at which Adams's two spheres differ only to that degree. So interpreted, the argument counts as a *continuity* argument because it relies on the assumption that if there are possible worlds that exhibit a range of degrees of a certain characteristic, and if the degrees in that range approach some limit, then there must be a possible world at which the limit is realized.

---

[14] It should be noted that the proof exploits the fact that *Mxy* includes as one of its conjuncts an existence claim for *x* and *y*, namely $(\exists z\ z{=}x\ \&\ \exists z\ z{=}y)$.

[15] Forrest, "The Identity of Indiscernibles," in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy (Fall 2006 Edition)*, URL = <http://plato.stanford.edu/archives/fall2006/entries/identity-indiscernible/>, at Section 3.

[16] Rodriguez-Pereyra, "The Bundle Theory is Compatible with Distinct but Indiscernible Particulars," *Analysis*, LXIV (2004): 72–81, at p. 74.

One consequence of our analysis of the argument of Passage II is that the argument in no way turns on an assumption that the space of possible worlds must be continuous in the way described above. A series of worlds with smaller and smaller differences between the spheres can serve to make the relevant causal independence claim seem more plausible and thus stimulate the intuition *that* the possibility of the indiscernibles-world follows from the possibility of a series of almost-indiscernibles worlds, but that is all that continuity can do. It is causal independence, not continuity, that gets us from the almost indiscernible to the indiscernible.

III. CAUSAL INDEPENDENCE AND THE ESSENTIALITY OF ORIGINS

Kripke makes his famous argument for the Essentiality of Origins in footnote 56 of *Naming and Necessity*:

> A principle suggested by [my] examples is: *If a material object has its origin from a certain hunk of matter, it could not have had its origin in any other matter.* Some qualifications might have to be stated (for example, the vagueness of the notion of hunk of matter leads to some problems), but in a large class of cases the principle is perhaps susceptible of something like proof, using the principle of the necessity of identity for particulars. Let '*B*' be a name (rigid designator) of a table, let '*A*' name the piece of wood from which it actually came. Let '*C*' name another piece of wood. Then suppose *B* were made from *A*, as in the actual world, but also another table *D* were simultaneously made from *C*. (We assume that there is no relation between *A* and *C* which makes the possibility of making a table from one dependent on the possibility of making a table from the other.) Now in this situation $B \neq D$; hence, even if *D* were made by itself, and no table were made from *A*, *D* would not be *B*. Strictly speaking, the 'proof' uses the necessity of distinctness, not of identity. However, the same types of considerations that can be used to establish the latter can be used to establish the former.... In any event, the argument applies only if the making of *D* from *C* does not affect the possibility of making *B* from *A*, and vice-versa (*op. cit.*, p. 114, note 56).

Kripke's argument in this passage is notoriously incomplete, but the reconstruction that Salmon (*op. cit.*, p. 214, note 10) reports Kripke as having endorsed, which Salmon christens the Fat Argument, contains the following premise:

> (III) For any possible tables x and x', and any possible hunks of matter y and y', if it is possible for table x to be originally constructed from hunk y while hunk y' does not overlap with hunk y, and it is also possible

> for table x' to be originally constructed from hunk y', then it is also possible that table x be originally constructed from hunk y and in addition table x' be originally constructed from hunk y'.[17]

To simplify the discussion, let us instantiate (III) with the names of Kripke's two possible hunks of matter a and c replacing the variables y and y', respectively, and where Txy means that table x is originally constructed from hunk of matter y, and Oxy states a relation between x and y which, as Kripke (*op. cit.*, p. 114, note 56) puts it, "makes the possibility of making a table from [x] dependent on the possibility of making a table from [y]." In what follows, for the sake of brevity, we will follow Salmon in taking Oxy to mean that hunk of matter x *overlaps* with hunk of matter y, but this is an oversimplification. Strictly speaking, Oxy amounts to a disjunction of.'x overlaps y' with other conditions. With this in mind, instantiating y and y' with a and c, respectively, yields the following:

(25) $\Box\,\forall x\,\Box\,\forall x'\ ((\Diamond(Txa\ \&\ {\sim}Oac)\ \&\ \Diamond Tx'c) \supset \Diamond(Txa\ \&\ Tx'c))$

As before, we assume quantified S5 as our underlying system of logical consequence. By S5, (25) is equivalent to the following:

(26) $\Box\,\forall x\,\Box\,\forall x'\ \Box((Txa\ \&\ {\sim}Oac\ \&\ \Diamond Tx'c) \supset \Diamond(Txa\ \&\ Tx'c))$

Consider, in addition, the following auxiliary premise, which states that the same possible table could not have been made originally from both of two nonoverlapping hunks of matter a and c:

(27) $\Box\,\forall x\,\Box({\sim}Oac \supset {\sim}\Diamond(Txa\ \&\ Txc))$

Premises (26) and (27) jointly entail the substantive essentialist conclusion (28) that no possible table x made originally from a could have been made originally from c:

(28) $\Box\,\forall x\,\Box((Txa\ \&\ {\sim}Oac) \supset {\sim}\Diamond Txc)$

Although (28) states a claim of origin essentialism, one might think that (26) does not and is instead as innocuous as (27). After all, (26) seems merely to specify that, given the absence of an interference-creating relation between a and c, making x from a is compossible with making x' from c provided that x is in fact made from a in a world in which making x' from c is at least possible. The fact that a and c do not overlap[18] seems to be the key condition: if a and c

---

[17] Salmon, *op. cit.*, p. 213. Note that in saying that x is originally constructed from hunk of matter y we mean that in the original construction of x, all of y was used, and y exhausts the matter that was used.

[18] Recall that *overlap* here does duty for an unspecified disjunction of interference-creating relationships.

had instead overlapped, compossibility would surely not obtain. This argument for the innocuousness of (26) is misleading, however, as Salmon argues. The reasoning from (26) and (27) to (28) relies only on the case of (26) where x and x' are instantiated by the same thing:

> In fact, this special instance of (III), with x and x' the same, is the only instance actually needed in the Fat Argument. Although this premise is slightly weaker than Kripke's essentialist thesis ... it hardly has any less essentialist import (*op. cit.*, pp. 213–14).

In other words, the only "part" of (26) that the Fat Argument depends on is this:

(29) $\Box\ \forall x\ \Box((Txa\ \&\ {\sim}Oac\ \&\ \Diamond Txc) \supset \Diamond(Txa\ \&\ Txc))$

And (29), Salmon argues, does have nontrivial essentialist import. Why did (26) then appear innocuous? Salmon writes as follows:

> One reason is that the bound variables 'x' and 'x'' occurring in the statement of (III) are distinct, and consequently the reader immediately thinks of instances where these two variables take on distinct entities as values. For *distinct* tables x and x', principle (III) is quite plausible, and indeed quite free of any nontrivial essentialist import (*op. cit.*, p. 214).

Why are (III), and, accordingly, (26), plausible in the case where $x \neq x'$, and why does what makes them plausible in that case not make them equally plausible in the case where $x = x'$? I will argue that causal independence is what makes the difference.

For the sake of simplicity, consider the result of instantiating (26) with possible tables b and d replacing the two variables x and x', respectively:

(30) $\Box((Tba\ \&\ {\sim}Oac\ \&\ \Diamond Tdc) \supset \Diamond(Tba\ \&\ Tdc))$

The claim of (30) is that, necessarily, if b is made originally from a, and if a does not overlap with c, and if making d originally from c is possible, then making b originally from a and making d originally from c are compossible. The antecedent of the conditional in (30) divides into two cases according to whether $b = d$:

(31) $\Box((Tba\ \&\ {\sim}Oac\ \&\ \Diamond Tdc\ \&\ b \neq d) \supset \Diamond(Tba\ \&\ Tdc))$

(32) $\Box((Tba\ \&\ {\sim}Oac\ \&\ \Diamond Tdc\ \&\ b = d) \supset \Diamond(Tba\ \&\ Tdc))$

What reason is there for thinking (31) is true? Descriptions by Kripke and others of the worlds in which Kripke's tables are made suggest that these worlds are altogether nomologically mundane, that is, they are worlds in which the actual world's laws of nature hold. Keeping this in mind, one can argue for (31) on the basis of considerations of causal independence using the following premise, *with the new understanding* *that Txy means that table x is constructed originally from hunk y in mundane circumstances, that is, in a world in which the actual world's laws of nature hold:*

(33) $\Box((Tba \mathbin{\&} {\sim}Oac \mathbin{\&} \Diamond Tdc \mathbin{\&} b \neq d) \supset N(Tba, Tdc))$

That is, (33) means that if table *b* is made originally from hunk of matter *a* in mundane circumstances and *a* does not overlap with hunk of matter *c* and making table *d* originally from *c* in mundane circumstances is possible, where tables *b* and *d* are distinct, then *b*'s being made from *a* in mundane circumstances is causally independent of whether *d* is made from *c* in mundane circumstances. (31) follows from (33) in an extension of quantified S5 for which the following holds:

(34) $N(Tba, Tdc), Tba, \Diamond Tdc \vdash \Diamond(Tba \mathbin{\&} Tdc)$

Since $\Diamond(Tba \mathbin{\&} Tdc)$ follows from *Tba* and *Tdc* in S5, (34) in turn holds for any system that extends S5 to include (22), one instance of which is the following:

(35) $N(Tba, Tdc), Tba, {\sim}Tdc, \Diamond Tdc \vdash \Diamond(Tba \mathbin{\&} Tdc)$

Thus, (31) follows validly from (33) given (22) and S5, and the inference is mediated by the causal independence of *Tba* from *Tdc*. If the laws of nature with which we are familiar did not make *Tba* causally independent of whether *Tdc* holds, there would be no reason to think *Tba* and *Tdc* were compossible.

Consider, on the other hand, the case of (32), in which $b = d$ appears. By the Necessity of Identity, S5, and the usual rules for identity, (32) is equivalent to the following:

(36) $\Box((Tba \mathbin{\&} {\sim}Oac \mathbin{\&} \Diamond Tbc) \supset \Diamond(Tba \mathbin{\&} Tbc))$

Our argument for (31) cannot be mutated into a sound argument for (36) because the counterpart of (33) in that argument would be the following, which seems not to be true:

(37) $\Box((Tba \mathbin{\&} {\sim}Oac \mathbin{\&} \Diamond Tbc) \supset N(Tba, Tbc))$

Suppose we are in a world *w* at which *b* is made originally from *a* in mundane circumstances but *b* could have been made originally from *c* in mundane circumstances, where *a* and *c* do not overlap. At such a world, making *b* originally from *a* can hardly be causally independent of whether *b* is made originally from *c*. This follows straightforwardly given our account (5) of causal independence. *Tba* is true at *w*, and we can suppose that ${\sim}Tbc$ is true at *w* by (27), since *Tba* and ${\sim}Oac$ are both true at *w*, so the conditional ${\sim}Tbc > Tba$ is true at *w*, by (6). But since *a* and *c* do not overlap, it is not the case that *b* would be made originally from *a* if *b* were made originally from *c*, that is, $Tbc > Tba$ is false at *w*. By (5), or by (18) if one prefers, it follows that $N(Tba,Tbc)$ is false at *w*. That is, at *w*, *b*'s being originally made from *a* is *not* causally independent of whether *b* is originally made from *c*. Thus (37) is false.

The question that motivated our discussion was this: Why are (III), and, accordingly, (26), plausible and innocuous in the case where $x \neq x'$, and why does what makes them plausible and innocuous in that case not make them equally plausible and equally innocuous in the case where $x = x'$? Our discussion of (31) and (32) has revealed the answer: in the case where $x \neq x'$, the compossibility of Txa and $Tx'c$ is explained by the causal independence of the former from the latter. This causal independence does not and cannot hold if $x = x'$, so if (III) and (26) hold in the case where $x = x'$, it is for some other reason.

IV. THE IDENTITY OF INDISCERNIBLES AND ESSENTIALISM

Della Rocca blames the influence of Black's and Adams's arguments against the PII on the New Theory of Reference and essentialist accounts of transworld identity:

> The philosophical climate must be right for such counterexamples to be taken seriously and it has been the misfortune of PII that other developments in philosophy have of late created an environment in which, for general theoretical reasons, such counterexamples to PII seem particularly compelling. To be more specific, I would say that the so-called new theory of reference and the standard essentialist account of transworld identity have a close affinity with the rejection of PII, even if this rejection does not stand in any direct entailment relation with these other theses (*op. cit.*, p. 481).

Two points are worth making in response to the above.

First, our discussion of the argument of Passage II shows that philosophers wanting to defend the PII against that argument can do so without sacrificing either the Necessity of Identity or the New Theory of Reference. A substantively different approach is available, namely the strategy of arguing against the metaphysical possibility of laws of nature L satisfying (23). I am not persuaded that this strategy would be effective, but, effective or not, it is an option whose existence was hidden by the 'surely' in Passage II.

Second, while I do not take Della Rocca's reference to "the standard essentialist account of transworld identity" to amount to an equation of nontrivial essentialism with the thesis of transworld identity, it is nevertheless worth emphasizing that such an equation would be mistaken and that nontrivial essentialism is most assuredly not involved in symmetrical universe counterexamples to the PII. Indeed, our discussion in this paper explains precisely *why* nontrivial essentialism is not involved in the argument of Passage II. That argument is free of nontrivial essentialist import for precisely the same reason that, when $x \neq x'$, (III) is free of such import, namely because causal independence implies compossibility, per (21) and (22), *precisely when no nontrivial essentialism is being invoked.* The argument of Passage II does depend on the Necessity of Identity, as Adams knew from the start, and examples such as the identification of water with H_2O suggest a close association between substantive, nontrivial essentialism and the Necessity of Identity. Our discussion is a reminder that the Necessity of Identity can have application even where no nontrivial essentialism has been assumed. The argument of Passage II is precisely such a case.

CHARLES B. CROSS

University of Georgia

COMMENTS AND CRITICISM

COLLECTIVE INTENTIONS AND GAME THEORY*

In 1988 I published a paper entitled "We-Intentions," written jointly with Kaarlo Miller.[1] The paper has been widely discussed in the literature and, unfortunately, often misunderstood.[2] In their recent paper that appeared in this JOURNAL, Natalie Gold and Robert Sugden criticize some philosophical accounts of collective (or joint) intention and present their own "team reasoning" account.[3] One of the criticized accounts is the mentioned Tuomela-Miller (TM) account. As I still basically hold on to the account given in the mentioned joint paper (although with considerable enrichments and improvements made in several later papers and books—see below), I will here briefly respond to the points that Gold and Sugden make.

I. WE-INTENTIONS

I will not discuss Gold and Sugden's (GS) own account of collective intentions at length. However, I will argue that it needs some changes and that with those changes it will get very close to my account. It is, however, worth pointing out that they are primarily investigating the kind of reasoning that supposedly leads to the formation of collective (here, equivalently, joint) intentions and they do not attempt to elucidate the detailed content of we-intentions—namely, the individual participants' "parts" of joint intentions (which is the main focus of the TM paper). My main concern here will be to show that Gold and Sugden's critical points against the TM account are based on a misunderstanding of my way of thinking about collective intentions and of some of the details in the paper. This is mainly due to their ignoring the explicitly mentioned presuppositions of the TM account—which

---

*I wish to thank Raul Hakli, Kaarlo Miller, and Maj Tuomela for comments.

[1] See Raimo Tuomela and Kaarlo Miller, "We-Intentions," *Philosophical Studies*, LIII (1988): 367–89. This paper is based on *A Theory of Social Action* (Boston: Reidel, 1984), but it improves on it concerning some details. It was written for presentation at the American Philosophical Association fall meeting in 1985.

[2] John Searle's flawed discussion of the Tuomela-Miller paper has been a central source of misunderstandings in the literature—see below and Searle, "Collective Intentions and Actions," in Philip Cohen, Jerry Morgan, and Martha Pollack, eds., *Intentions in Communication* (Cambridge: MIT, 1990), pp. 401–15.

[3] See Gold and Sugden, "Collective Intentions and Team Agency," this JOURNAL CIV, 3 (March 2007): 109–37.

0022-362X/09/0605/292–300

in effect deny the possibility of treating the participants of joint intention as individually rational agents playing strategically against each other (which is what Gold and Sugden assume).

Suppose some agents jointly intend to perform a joint action or to achieve a goal together—for example, they jointly intend to build a bridge together, to sing a song together, or write a joint paper. In purest cases they can be taken, in effect, to function in the "group-mode" (or "we-mode," in my terminology) and to accept the statement 'We will do *X*' as expressing both their joint intention (when read collectively) and their individual we-intentions (when read distributively) in an intuitive, preanalytic sense.[4] As is said in the TM paper, what is being analyzed and elucidated in the paper is an irreducible notion of we-intention (TM 368). What exactly the presupposed preanalytic notion of we-intention (and joint intention) is, is left somewhat open in the paper, which mainly purports to elucidate what having a we-intention amounts to, given the starting point of the intuitive notion. A minimal intuitive idea here is that the participants are supposed to function *as a group* or as one agent and appropriately to coordinate or bind together their activities both in their reasoning and acting as group members. This idea seems to be shared by the GS account.[5] However, the GS account does not really analyze collective intentions but only the process leading to the rational formation of them (see below).

In the TM account the participants' joint intention to perform an action *X* entails that the participants we-intend to perform *X* together. Thus, if John and Mary jointly intend to perform an action *X* (for example, to lift a table) together, this joint intention is something that they have between themselves concerning the joint action in question. Here both John and Mary are taken to we-intend X, their joint lifting of the table.[6] The TM paper focuses on we-intentions rather than joint intentions and takes X to be a joint action type, namely, roughly a jointly tokenable action structure based on we-thinking and involving part actions for at least two participants.[7] In a choice situation of the kind Gold and Sugden deal with, joint tokening of an action type minimally involves that the participants choose as a unit, their choices thus being in a sense "glued" together. This assumption is also shared by the GS account precisely because these authors are analyzing team reasoning.

---

[4] As I have emphasized especially in my later works, this joint intention provides a *group reason* for them to intend to perform their parts in the joint action *X* that is purported to satisfy the joint intention and to the performance of which the participants are *collectively committed* (because of their jointly intending *X*); see especially Tuomela, *The Philosophy of Sociality: The Shared Point of View* (New York: Oxford, 2007). A joint intention in the sense meant here is a we-mode joint intention that satisfies also the strong *collectivity condition* according to which, on the basis of the participants collectively having accepted the joint intention for their group, it is necessarily the case that when the intention is satisfied for any of the group members it is satisfied for all of them. (This is a kind of strong "being in the same boat" type of condition that is not to be confused with the participants just happening to share the same goal.)

When GS speak about "collective intention" they seem to mean what I mean by the term "joint intention." Sometimes they say "collective intentionality" in a similar context (for example, on GS 111). This, however, is not an acceptable term here because not only joint intentions but also, for example, joint wants, beliefs, emotions, and actions express collective intentionality (or "aboutness").

[5] In the case of a two-person Hi-Lo situation, for instance, Gold and Sugden deal with a team of agents who combine their actions and collectively rationally choose (high, high). This kind of team action is part of what I mean by acting as a group in the present situation.

The TM assumption that the participants accept the intention expression 'We will do X' (or 'We will do X together') is to be understood as a kind of presupposition of a more detailed account of their having the we-intention to perform X together (TM 368–69). This of course goes together with the view that an irreducible notion of we-intention is being elucidated. The TM account also emphasizes and presupposes that the participants should be disposed to make relevant practical inferences, such as expressed by the following simple schemas ($W1$) and ($W2$) (see TM 268):[8]

($W1$) (i) We will do X.
(ii) I am one of us, namely, the group G.

[Therefore,]

(iii) I will do my part of X.

Here the we-intention qua a "slice" of the participants' joint intention to perform X gives the agent ("me") a group reason or "we-reason" to intend to perform (and to perform) his part of X.

---

[6] From a logical point of view we may treat intentions as predicates. Letting 'JI' stand for joint intention and 'WI' for we-intention, we have roughly this equivalence in the case of two agents A and B and an action X: $JI(A,B,X) \leftrightarrow WI(A,X)$ & $WI(B,X)$ & $MB(WI(A,X)$ & $WI(B,X))$, where 'MB' stands for mutual belief. However, this logical formulation is somewhat misleading (although truth-functionally correct), as it ignores the fact that a we-intention *intentionally* (in the content of a we-intention) presupposes the existence of the joint intention (see Tuomela, *The Philosophy of Sociality*, chapter 4, for discussion of this circularity problem). An intention attributed to a group G, namely, $I(G, X)$ is to be distinguished from the joint intention of its members in general, although in the case of simple unstructured, egalitarian groups, the group intends if and only if its members do.

[7] See Tuomela, *A Theory of Social Action*, chapter 5, and for a newer account with an emphasis on functioning as a group, see chapters 5 and 6 of Tuomela, *The Philosophy of Sociality*.

[8] For my later formulations, see, for example, the somewhat more general ones in Tuomela, *The Philosophy of Sociality*, chapter 4.

(*W2*) (i) We will do X.
(ii) A is one of us.
(iii) Our doing X requires that A does his part of X.
(iv) Unless I do Y (for example, teach A to do something [required of him for X]), A cannot do his part.

[Therefore,]

(v) I will do Y.

Given these presuppositions, essentially the following summary elucidation of we-intentions is given in the 1988 paper (TM 375):

(*WI*) A member A_i of a collective G *we-intends* to do X if and only if

(i) A_i intends to do his part of X;
(ii) A_i has a belief to the effect that the joint action opportunities for X will obtain, especially that at least a sufficient number of the full-fledged and adequately informed members of G, as required for the performance of X, will (or at least probably will) do their parts of X;
(iii) A_i believes that there is (or will be) a mutual belief among the participants of G to the effect that the joint action opportunities for X will obtain;

In the text of the paper there are some further qualifications such as that the participants must do their parts intentionally as their parts of X and that they are supposed to we-intend "with the purpose of the action X coming about" (TM 377). Furthermore, the beliefs attributed to the agents are supposed to have some causal impact on the participants' intentions to perform their parts—this feature does not apply to the game-theoretical approach of Gold and Sugden, causality playing no role in that framework.[9]

Although the summary account (*WI*) has the form of an explicit definition, the full account given in the paper does rely on the satisfaction of the mentioned two presuppositions. Denoting by *PS1* the presupposition of the participants' mutual acceptance of 'We will do X together' and by *PS2* the practical inference schemas in accordance with which the participants are assumed to be disposed to infer, the full account schematically becomes in the two-person case: *Given that the presuppositions PS1 and PS2 hold true, the account (WI) states necessary and jointly sufficient conditions for a participant to we-intend to perform the joint action X.*

---

[9] In my later work a somewhat enriched formulation has been used—for example, in Tuomela, *The Importance of Us: A Philosophical Study of Basic Social Notions* (Stanford: University Press, 1995) and in *The Philosophy of Sociality*. Let me here reproduce the formulation in chapter 4 of the latter work:

(*WI*) A member A_i of a collective g we-intends to do X if and only if

(i) A_i intends to do his part of X (as his part of X);
(ii) A_i has a belief to the effect that the joint action opportunities for an intentional performance of X will obtain (or at least probably will obtain), especially that a right number of the full-fledged and adequately informed members of g, as required for the performance of X, will (or at least probably will) perform their parts of X, which under normal conditions will result in an intentional joint performance of X by the participants;

I will not here discuss the details of the TM account and thus not, for example, John Searle's criticism of its alleged circularity.[10] However, I will consider Searle's alleged countererexample to the account partly because it bears similarity to the examples by Gold and Sugden. In Searle's alleged counterexample to the TM account, we are invited to suppose that "a group of businessmen are all educated at a business school where they learn Adam Smith's theory of the hidden hand. Each comes to believe that he can best help humanity by pursuing his own selfish interest, and they each form a separate intention to this effect; that is, each has an intention he would express as 'I intend to do my part toward helping humanity by pursuing my own selfish interest and not cooperating with anybody'. Let us also assume that the members of the group have a mutual belief to the effect that each intends to help humanity by pursuing his own selfish interests and that these intentions will probably be carried out with success."[11] Searle now claims that this example refutes the TM account. However, he does not deal with the full TM account at all but only a part of it. The main reason is that neither presupposition *PS1* nor *PS2* is satisfied here, and furthermore, the businessmen are not pursuing parts of any joint action as their parts of it. Searle's example thus has no bite against the TM account.

II. GOLD AND SUGDEN ON WE-INTENTIONS

Let me now consider Gold and Sugden's critical points. One of their central theses against my account is that "it seems to treat every Nash

---

(iii) A_i believes that there is (or will be) a mutual belief among the participating members of g (or at least among those participants who perform their parts of X intentionally as their parts of X there is or will be a mutual belief) to the effect that the joint action opportunities for an intentional performance of X will obtain (or at least probably will obtain);

(iv) (i) in part because of (ii) and (iii).

[10] See the discussion in Searle, "Collective Intentions and Actions," and my responses in, for example, Tuomela, *The Importance of Us*, chapter 3, on Searle's interpretation of the TM paper, and in *The Philosophy of Sociality*, chapter 4, on the circularity problem.

[11] Searle, "Collective Intentions and Actions," p. 404.

equilibrium as a case of collective intentionality" and, accordingly, for example, in the case of a Prisoner's Dilemma "the Nash equilibrium behavior seems to fall under Tuomela and Miller's analysis of collective intentions" (GS 111–12).[12] While the authors discuss both the dove-hawk game and the Prisoner's Dilemma, they take the latter to make the critique against the TM account "even more acute."[13] To concentrate on the Prisoner's Dilemma consider the authors' example:

		*P*2	
		cooperate	defect
*P*1	cooperate	3,3	1,4
	defect	4,1	2,2

The interpretation they consider is a "take-some game" with the structure of a Prisoner's Dilemma. Here the exploitation of the resource amounts to choosing (defect, defect), which is the unique Nash equilibrium of the game, while the case of conserving the resource, namely, (cooperate, cooperate), does indicate the presence of a collective intention (in my language, joint intention). Gold and Sugden assume that the we-intending participants function individually rationally in the standard game-theoretic sense and form the intention to defect. However, this assumption is not entailed by the TM account and is even inconsistent with it. In the TM account the individual players are assumed not to be disposed to play against each other at all in the standard game-theoretic sense (thus concerning their private, I-mode utilities) and thus not to search for a best-reply strategy (action) against the other players' strategies (actions). Rather, the participants of a joint intention and action are assumed to be engaged in we-mode thinking and action for their group. They can accordingly form the collectively rational intention to conserve the resource. Yet, as conceptual possibility, they can also each form the *collectively nonrational* we-intention to defect and thus to exploit the resource that leads to a Pareto-suboptimal outcome.[14]

---

[12] Gold and Sugden also claim (GS 111) that the TM account "reduces we-intentions to individual intentions and a network of mutual beliefs." However, this is not only denied in the TM paper (TM 367) but it is quite clearly the case that the presuppositions *PS*1 and *PS*2 indicate the central reason why the TM account is not reductionistic. As I emphasize in my later works (most recently in Tuomela, *The Philosophy of Sociality*, chapter 4), the intentions under consideration are we-mode intentions, and this accordingly applies also to the participants' intentions to perform their parts of *X*. They might in addition also privately (namely, in the I-mode) intend to perform their parts—and would thus intend (given the usual presuppositions for I-mode intentions hold true) to continue to act relevantly even if the other participants had given up the joint project (for example, if you give up our project to write a paper jointly I might still write the paper by myself).

[13] In the authors' Hawk-Dove game there are two equilibria, namely, (dove, hawk) and (hawk, dove). Contrary to these authors' claim, neither of these equilibria are required by the TM account to be we-intended. As I will show below, the TM account actually does not entail that the object of a we-intention must be an equilibrium at all.

As said, the TM account does not entail either of these two possibilities as such, but when *collectively rational* joint intention and we-intention formation is at stake, it recommends the intention to cooperate (the we-intention to choose *C*), which indeed will be a dominant action for the participants operating in the we-mode as proper group members. In that case they adopt the group's point of view and take the cooperative alternative to be the one a rational group should choose, and when acting as proper group members they do choose *C*.

III. TEAM REASONING AND WE-INTENTIONS

As to the constructive part of the GS account, I accept much of it as right but have still a few central criticisms against it. The first one is that their account at best gives a sufficient condition for collective intentions, because of their strict adherence to the game-theoretic framework. In particular, the only collective goal that they consider and seem to take to be possible in their account is maximization of collective utility. This is a highly general goal that seems not to be applicable to all contexts where people intend collectively and cooperatively—think of walking together or singing a song together. Such cases basically are not cases for the exercise of "Machiavellian intelligence" and strategic thinking. In cooperative we-mode contexts the idea of the participants' strategic maximization of collective utility in strategic interaction need not be at stake and such a view is based on the wrong perspective on the situation. Furthermore, the commonsense goals that people collectively intentionally adopt and purport to achieve are typically more specific goals than maximizing expected utility (even if a theoretician may wish to view them in that way).

According to the GS account, the right way to go about analyzing collective intentions is not through the properties of intentions as mental states but in terms of the mental reasoning processes by which these intentions are formed (GS 110). My own account goes in terms of the we-mode/I-mode distinction. These modes indeed do involve the requirement of appropriate reasoning (recall the schemas (*W*1) and (*W*2) expressing we-mode or group-reasoning). However, notice that there seem to be spontaneously formed intentions (for example, relying on suitable background knowledge and patterns of intention formation) in which no actual reasoning takes place (while such reasoning could have taken place). The we-mode or group-mode approach takes this into account because it does not require actual reasoning in all cases, but the GS account seems to fail on this count at least as it has been formulated in the paper.

---

[14] Gold and Sugden also claim that the exploitation of the resource is not a joint action. But this is false, as any joint outcome can represent the kind of joint action type that the TM account is concerned with, and the joint intention (shared we-intention) to bring about (defect, defect) is a conceptual (although not a collectively rational) possibility according to our account.

As to the practical reasoning schemas of the GS account, consider, for example, its central *Schema 4* that concerns the individual participant's reasoning (GS 126):

(1) I am a member of *S*.
(2) It is common knowledge in *S* that each member of *S* identifies with *S*.
(3) It is common knowledge in *S* that each member of *S* wants the value of *U* to be maximized.
(4) It is common knowledge in *S* that *A* uniquely maximizes *U*.

Therefore,

(5) I should choose my component of *A*.

Here *S* is just a set of individuals (thus not necessarily a social group), *A* is a profile of acts for the members of *S* and *U* is a utility function. Identification with *S* is said to take place if the member in question conceives of *G* as a unit of agency. This seems to me to be only a necessary condition for identification (whatever identification means, in my view it should amount to functioning as a proper member of *S*) under a normal, *nonconative* interpretation of 'conceives'. The analogous group version basically aggregates the individual reasoning processes (see GS 125).[15]

My main critical points against Schema 4 are the following, supposing that its conclusion must be true whenever its premises are. If thus the above GS Schema 4 is taken to represent logically valid practical reasoning, it seems that mere wanting in (3) does not suffice for bringing about an all-things-considered "should" in (5), which seems needed for an intention (in the present context a proper conclusion of practical reasoning).[16] This is because the agents might have stronger incompatible wants, and in such a case it seems that they need not be required to choose their components of the profile *A* (but do something else, irrespective of (4)). What the schema does succeed in is arriving at a "technical" conclusion with a "should" whose normative force is left unspecified and which as such falls short of entailing that a rational agent will form an intention. However, it is said in the GS paper that team reasoning, as represented by Schema 4, results in the formation of intention (GS 126). Also the weaker point is made that if (5) indeed comes about within the context of Schema 4, then the team-reasoning participant will here have the reason to form the intention to choose his component of *A* (GS 126). But if we rely on the point that a mere want does not lead to an all-things-considered "should" in (5), we do not always arrive even at a well-founded reason to form the intention and still less to the kind of all-things-considered reason that the rational formation of an intention requires. The intention to act should thus be formed on the all-things-considered want, and a resulting all-things-considered "should" should be involved in this situation.

---

[15] The conclusion of the group version is: Each of us should choose her component of *A*.

[16] However, one of the authors prefers, after all, to deal with "success promotion" rather than logical validity in the case of schemas like the present one. However, success promotion also seems to require that the want in question be an intention, if the agent in question is rational (which the GS account requires), for otherwise the performance of the successful action component of *A* need not follow: A want, even if unique, need not always yield action in contrast to an intention which does entail setting oneself to act (barring akrasia). So even within the success promotion view the conclusion of my argument in the text seems to stand.

To end and to emphasize the similarity of the GS account and my approach, let me quote a sentence from the GS paper: "It is an implication of our analysis that a given pattern of behavior can be intended either individually or collectively, depending on the reasoning which led to it" (GS 128). According to my we-mode account, a piece of joint behavior can be intended either individually (namely, in the I-mode) or collectively (namely, in the we-mode), depending on which of these modes is active in the situation which depends on whether they act as group members in the full sense or as private persons.[17] This may but need not depend on an actual antecedent process of a respective kind of reasoning.

RAIMO TUOMELA

University of Helsinki and University of Munich

---

[17] See Tuomela, *The Philosophy of Sociality*, especially chapters 4–5, 7.

BOOK REVIEWS

The Structure of Objects. KATHRIN KOSLICKI. New York: Oxford University Press 2008. xix + 288 p. Cloth $ 90.00.

Kathrin Koslicki's *The Structure of Objects* is a study of parts and wholes. Koslicki critically discusses, often in great detail, a number of philosophers on this topic, from Plato and Aristotle to David Lewis. Her discussion is interesting not only for its own sake, but also because it prepares the ground for her original theory of composite objects. And it is Koslicki's theory that will be my focus here.

I

Koslicki's theory consists of two theses. The first thesis is an answer to the "special composition question." That question asks what it takes for some *x*s to compose an object *O*. In one way or another, most recent answers to this question are revisionary. Consider, for example, Koslicki's *bête noire*: "standard mereology." Standard mereology says, among other things, that any existing entities compose an object; thus it implies that your left ear and your neighbor's dog compose an object.

Koslicki wants to defend a nonrevisionary answer to the special composition question, an answer that is consistent with the existence of all and only the composite objects endorsed by common sense and by natural science. Her answer is that the *x*s compose an object *O* of kind *K* if and only if the *x*s satisfy the formal constraints dictated by the structure associated with objects of kind *K* (187–88). And Koslicki's answer is, as she intends, consistent with a commonsensical, scientifically informed, ontology.

But her answer does not deliver such an ontology, at least not on its own. After all, even standard mereologists, who believe in non-commonsensical sums of ears and dogs, can accept her answer. For standard mereologists can claim that *arbitrary sum* is a kind of object whose structure dictates that the parts of any arbitrary sum must satisfy the "formal constraint" of existing.[1] And standard mereologists can add that, for every other kind of object, the parts of any object of that kind must satisfy the further formal constraints dictated by the structure associated with that kind of object. They will certainly say, for example,

---

[1] Koslicki has a robust notion of *kinds* (200–34). So she might argue that *arbitrary sum* is not a genuine kind. But to do so is to go beyond her answer to the special composition question. And so that answer, by itself, does not rule out standard mereology. And that is my only point here.

0022-362X/09/0605/301–07

that the *x*s compose a motorcycle if and only if the *x*s are arranged in a particular way.

So Koslicki's answer to the special composition question does not rule out standard mereology. So her answer does not constitute an objection to standard mereology, at least not on its own. But Koslicki does have an objection to standard mereology. Here is one version of that objection. Suppose that a motorcycle is composed of certain atoms. Suppose that those atoms existed two hundred years ago, but were widely scattered. According to Koslicki, standard mereology implies that that motorcycle itself existed two hundred years ago, as those scattered atoms. But, Koslicki would object, no entity that is currently a motorcycle existed two hundred years ago, not even as scattered atoms. So standard mereology is false (45).[2]

Koslicki is right that standard mereology says, among other things, that if the *x*s compose an object *O*, then whenever the *x*s existed, they composed *O* (and so *O* existed then). But standard mereology does not imply that any motorcycle existed two hundred years ago, not even if the atoms that now compose it existed then. At least, *standard* standard mereology does not imply this. For most standard mereologists—such as David Lewis—are perdurantists. And perdurantists say that a motorcycle is composed not of entire atoms, but rather of the *temporal parts* of atoms. They will add that no motorcycle was ever composed of the two-hundred-years-ago temporal parts of any atoms, not even if the atoms now composing a motorcycle existed two hundred years ago. So nothing in perdurantist standard mereology implies that a motorcycle existed two hundred years ago, not even as scattered atoms. So most standard mereologists are untouched by Koslicki's principal objection to standard mereology.

II

As we have seen, the first thesis of Koslicki's theory—her answer to the special composition question—involves structure. So does the second thesis. Let the *K*-structure dictate the formal constraints that parts must satisfy in order to compose an object of kind *K*. Her second thesis says that if an object is a member of kind *K*, then that object has the *K*-structure *as a part*. I think that this is the most exciting claim of the book. It is so exciting that it may strike some contemporary metaphysicians as beyond the pale.[3] But Koslicki shows that Plato and Aristotle endorsed claims more or less like her second thesis (chapters 5 and 6). This pedigree makes her second thesis worth taking seriously, as do Koslicki's interesting arguments for that thesis.

---

[2] Koslicki sometimes puts this objection in a misleading way: "Standard mereology cannot tell the difference between the motorcycle in running condition and the heap of disassembled parts" (4; see also p. 1). Compare: many of us—including Koslicki—think that someone who is now a great-grandfather existed years ago, as an infant child; but it would be misleading, at best, to charge that we thereby "cannot tell the difference" between an elderly gentleman and a baby.

Koslicki's first argument (180–81) begins with some familiar claims about a statue and a lump of clay. These claims are: the statue and the lump are in the same place at the same time; but the statue is not identical with the lump; instead, the statue is constituted by the lump. From these familiar claims, Koslicki infers that the lump is a part of the statue. Because the statue is not identical with the lump, she even infers that the lump is a *proper* part of the statue.

Enter the Weak Supplementation Principle (WSP): if *x* is a proper part of *y*, then there is some *z* that is a proper part of *y*, and that shares no parts with *x*. Given that the lump is a proper part of the statue, WSP tells us that the statue has a proper part that shares no parts with the lump. But the statue and the lump have all the same material parts.[4] So, given WSP, the statue must have a nonmaterial part that the lump lacks. That extra part, says Koslicki, is the statue's *structure*. Thus her first argument for the claim that an object has its structure as a part.

WSP is essential to Koslicki's first argument. But Koslicki does not argue for WSP.[5] Nor is WSP obviously true. In fact, here is a counterexample to WSP. Suppose that the statue is constituted by (but not identical with) the lump. And suppose, further, that this implies that the lump is a proper part of the statue. Finally, add that each of the statue's parts shares some part or other with the lump. All of this together implies that WSP is false.

---

[3] This is exciting. Of course, this would not be exciting if *having a structure as a part* were the same thing as *having a structure*. (For it is not exciting that a member of kind *K* has a *K*-structure.) So in order to appreciate Koslicki's second thesis, we must see that *having a structure as a part* is one thing, and *having a structure* is something else.

Here is one way to see this. Suppose, for *reductio*, that *having a structure* is the same thing as *having that structure as a part*; a cell in my finger has a certain structure; therefore (given our assumption for *reductio*) that cell has that structure as a part; that cell is a part of me, and so its parts are my parts; so that cell's structure is a part of me; and so (given our assumption for *reductio*) I have the same structure as that cell. RAA.

[4] More carefully, they have the same material parts *at some level of decomposition*. For example, they presumably have all the same molecules as parts. But the statue may have some material parts that the lump lacks. For example, the statue may have a head as a part, but the lump not. (The lump may have a head-shaped lump as a part.) Even so, the statue's head will share some parts—such as certain molecules—with the lump (and with the lump's head-shaped part). In what follows, I shall sometimes say that the statue and the lump "have the same material parts"; this is shorthand for saying that they have the same material parts at some level of decomposition.

[5] Koslicki takes WSP to be "partially constitutive of the meaning of 'is a proper part of'" (180). But I think that the meaning of 'is a proper part of' is entirely constituted by: (a) is a part of; and (b) is not identical with.

This "counterexample" to WSP should seem familiar. It is, of course, the example that drives Koslicki's first argument for the statue's having its structure as a part. So Koslicki cannot resist this counterexample by denying (as others might) that the lump is a proper part of the statue. I think that she must resist by saying something like: "WSP is more plausible than is the claim that each of the statue's parts shares some part or other with the lump." For what it is worth, I would judge the plausibilities differently.

And I have a more serious objection. This objection begins with a big object of kind *K*. Suppose that that object is constituted by a numerically distinct object of kind *K**. Suppose that this implies that that *K**-object is a proper part of the big *K*-object. Let us add that the *K**-object has a part that is itself of kind *K*, a small *K*-object. Suppose that the small *K*-object has the *K*-structure as a part. By the transitivity of parthood, the *K**-object has the *K*-structure as a part. WSP tells us that the big *K*-object must have a part that shares no parts with the *K**-object. That part cannot be the *K*-structure itself, since—as we have just seen—the *K**-object has the *K*-structure as a part. But—so Koslicki's first argument seems to imply—that part is the *K*-structure. Contradiction.

The contradiction just generated has two sources. The first source is the reasoning found in Koslicki's first argument for the claim that an object has its structure as a part. Obviously, Koslicki cannot reject this first source. The second source is my example that begins with a big *K*-object. So Koslicki must object to something or other in my example.

Koslicki herself—in the course of considering an objection different from the one I am raising here—discusses an example of just this sort: A big statue constituted by a heap of trash, a heap which includes, among other items, a little statue (256). Koslicki does not object to this example by denying that statues exist, or by denying that constitution is a relation that holds between distinct entities. Nor should she. After all, either denial would undermine the reasoning in her first argument.

Nor do I think that she should deny that heaps of trash exist (even though she entertains this denial on pp. 258–59). This is partly because eliminating heaps of trash alone would be a mere stopgap measure; we could come up with a less trashy example along these same lines.[6] Moreover, her central arguments require that statues made of clay are constituted by something—and why should statues made of trash be any different? Finally, she insists that a theory of parts and wholes should not lead to a revisionary account of what kinds of entities exist (171). To eliminate heaps of trash only because they create trouble for her theory of parts and wholes would make her the very sort of revisionary mereologist she sets out to oppose.

---

[6] Here is just one: a large ice cube constituted by a mass of (frozen) water composed of smaller ice cubes.

Koslicki could deny that the big statue's structure is numerically identical with the small statue's structure. For she could deny that structures are universals. And then she could endorse trope theory. This would allow her to say that the big statute has a part—namely, the trope that is that big statue's structure—that the heap of trash does not have. She can say this even though the heap of trash has, as a part, the trope that is the little statue's structure. And she can say this even though the two statues have the same structure, since on trope theory their having the same structure implies only that the structure of one is exactly similar to the structure of the other.[7]

Embracing trope theory seems to me the best way for Koslicki to object to the example that, I argued above, leads to contradiction when combined with her reasoning. But this embrace is not without costs for Koslicki, not even if trope theory is free. For trope theorists say that all of an object's properties are parts of that object. So they say that an object's structure—assume its structure is a property—is itself a part of that object. (Koslicki is open to the idea that a structure is a property; see p. 254.) So they agree with Koslicki's claim that an object's structure is a part of that object. Thus her claim turns out to be just one predictable upshot of trope theory, as opposed to a new account of the nature of objects, or a bold hypothesis about how an object's structure differs from its other properties. In this way, I think that trope theory makes Koslicki's claim less exciting than it originally seemed to be.

III

The statue and the lump are supposed to differ in persistence conditions. Presumably, their differing in this way must be grounded by their differing in some other way. But it seems that there is no other way in which they differ. After all, the statue and the lump have the same material parts and are in the same environment. This is the "grounding problem" for coincident objects.

Koslicki's second argument for the claim that an object has its structure as a part is that that claim solves the grounding problem. For suppose that claim is true. Then the statue differs from the lump by having a part—its structure—that the lump does not have. And that difference, says Koslicki, can ground their difference in persistence conditions (181–83).

---

[7] Koslicki entertains trope theory, and raises objections to it (257–58). She also entertains the idea that the big statue and the little statue do not have the same structure at all (257–58). But this latter idea is inconsistent with her view that members of the same kind have the same structure, at least given her assumption (in the central arguments for her theory) that *statue* is a genuine kind.

Suppose one says that the statue's differing in persistence conditions from the lump is grounded by the statue's, but not the lump's, being a member of the kind *statue*. Arguably, this "solution" merely moves the bubble to a new place under the rug. For we are now left wondering what grounds the postulated difference in kind membership. And it looks like nothing can. For the statue and the lump have the same material parts and are in the same environment.

This bubble-moving "solution" can be made to look a lot like Koslicki's solution. For suppose that the fundamental difference (which grounds every other difference) between the statue and the lump is that the statue exemplifies *being a statue*, but the lump does not. And suppose that trope theory is true, and that, as a result, an object has a property if and only if that object has that property as a part. Then the fundamental difference between the statue and the lump is that the statue has, but the lump lacks, an abstract part: *being a statue*. This looks like Koslicki's solution, assuming that a structure is a type of property. (Indeed, this *is* Koslicki's solution if her overall approach incorporates trope theory—recall the closing of the previous section.) And this makes me conclude that Koslicki's solution is but a bubble mover.

Here is another route to that same conclusion. Consider the lump's material parts. They are numerically identical with the statue's material parts. And so, given the indiscernibility of identicals, the lump's material parts are interrelated exactly like the statue's material parts are interrelated; and so the lump's parts satisfy all the same formal constraints that the statue's parts satisfy. All of this leaves us wondering what explains why the statue has the relevant structure as a part, but the lump does not. More to the point, it seems that nothing could explain this difference. And so it looks like Koslicki's solution to the grounding problem merely exchanges one objectionable brute difference for another.[8]

If the lump has the statue's structure as a part, then both of Koslicki's arguments for the claim that an object has its structure as a part fail. So Koslicki must say that the lump does not have the statue's structure as a part. And so Koslicki must deny that an object's having a certain structure as a part supervenes on how that object's material parts are interrelated. But even setting the grounding problem aside, I find this denial implausible. At least, if I were defending Koslicki's claim that an object has its structure as a part, I would insist that its having that structure as a part supervenes on how that object's material parts are interrelated. Put otherwise, I would insist that an object's having that structure as a part supervenes on that object's structure.

---

[8] Standard mereologists say that if there are some *x*s and a structure *S*, then—simply as a result of the unrestricted nature of composition—there is an object composed of just the *x*s, and another composed of just the *x*s and *S*. So, they will say, there is no need to explain why the latter object has *S* as a part, but the former does not, and so nothing objectionable about lacking such an explanation. Koslicki is no standard mereologist. She denies that composition is unrestricted. And so she should expect there to be an explanation of why one thing (like a statue) has another thing (like a structure) as a part.

IV

Suppose that a structure is a property. In particular, suppose that the structure associated with the kind *statue* is the property of *being structured like a statue.* Now suppose that a lump—in virtue of how its material parts are interrelated—exemplifies that property. Add that there is therefore the "event," or the Armstrongian "state of affairs," of that lump's exemplifying that property. According to the lore on events, that event has two principal parts. Its principal parts are the lump, and the property of *being structured like a statue.* (The event's other parts are all parts of these two principal parts.)

Dwell upon the event of the lump's exemplifying *being structured like a statue.* You will come to see that that event seems to be exactly like a statue, at least given Koslicki's account of statues. That is, Koslicki's account seems to be aptly described as the claim that a statue is an event, an event whose principal parts are a lump and the property of *being structured like a statue.*[9]

Some philosophers deny that numerically distinct statues and lumps can occupy the same place at the same time. But these philosophers do not deny that a lump can be in the same place at the same time as the event of that lump's exemplifying *being structured like a statue.* Moreover, no one even raises an eyebrow at the claim that that lump's persistence conditions differ from those of that event. Nor are we likely to object to the claim that that event has the property of *being structured like a statue* as a part, but that that lump does not—and this despite our knowing that the material parts of that lump are (identical with and so) interrelated exactly like those of that event.

So suppose that Koslicki's statue really is an event. Then some of my objections above might not indicate real problems for her account of statues and other objects. Or perhaps, instead, those objections indicate problems not just for her account, but also for the claim that there are events. I do not know which it is.

TRENTON MERRICKS

University of Virginia

---

[9] Koslicki's view also seems to imply that the lump is an event. More generally, her view seems to imply that all material objects are events, except for unstructured objects at the bottom of the compositional hierarchy. Koslicki is explicitly agnostic about whether there are unstructured objects (186–88).

MEMO TO AUTHORS

The Journal of Philosophy is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally publish to no more than 7,500 words, or 18 *Journal* pages. A more precise way of determining length is by characters (including spaces): 2,850 characters per *Journal* page, or 51,300 characters per *Journal* article of 18 pages.

The editors request that three copies (or an electronic version) of all manuscripts submitted be prepared according to the following guidelines:

(1) All material, including the quotations and footnotes, must be typed, double-spaced, and one-sided, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into footnotes.
(3) Footnotes should be complete, according to *Journal* style (see any issue). Journal citations should include: name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include: publisher, U.S. city of publication, and date. Footnotes should be gathered at the end of the article.
(4) Quotations should be checked by the author; capitalization, italics, spelling, and the like should match the original.
(5) If there are diagrams, they should be provided camera-ready by the author. If there are special symbols, they should be widely spaced and clearly marked. Standard logic symbols are available, and letters of the Greek and script alphabet.
(6) No underlining; please limit use of italics to technical terms (when they are introduced) and to variables.
(7) Minimize use of contractions.

The editors are willing to read and evaluate a manuscript that does not meet these guidelines, provided the author agrees to meet them before submitting a final manuscript. The editors may ask for revision of excessively long manuscripts, and may require reformatting of accepted manuscripts. They also request an electronic version with the final hard copy.

The Journal of Philosophy does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere. While we do not publish manuscripts that mainly comment on articles published elsewhere, we do have a "Comments and Criticism" section in which short responses to *Journal* articles may be published.

Manuscripts may be sent to: John Smylie, Managing Editor, *The Journal of Philosophy*, 709 Philosophy Hall, 1150 Amsterdam Avenue, Mail Code 4972, New York, NY 10027, Email submissions@journalofphilosophy.org.

THE JOURNAL OF
PHILOSOPHY

Special Issues Available

Epistemic Norms - September and October 2008
Parts and Wholes - December 2006
On Bernard Williams - June 2003
Causation - April 2000
Jürgen Habermas and John Rawls - March 1995

The Dewey Lectures:
Michael Dummett - *Truth and the Past,* 2003
Hilary Putnam - *Sense, Nonsense, and the Senses:*
An Inquiry into the Powers of the Human Mind, 1994
Donald Davidson - *The Structure and Nature of Truth,* 1990
Amartya Sen - *Well-being, Agency, and Freedom,* 1985
John Rawls - *Kantian Constructivism in Moral Theory,* 1980
Ernest Nagel - *Teleology Revisited,* 1977
W.V. Quine - *Ontological Relativity,* 1968

The Woodbridge Lectures:
John McDowell - *Having the World in View:*
Sellers, Kant, and Intentionality, 1997

Subscriptions
Individuals $45.00
Students $20.00
Libraries and Institutions $100.00
Print and Online for Libraries and Institutions $180.00
Postage outside the United States $20.00

www.journalofphilosophy.org

New from Stanford University Press

"What Is an Apparatus?" and Other Essays

GIORGIO AGAMBEN Translated by
DAVID KISHIK and STEFAN PEDATELLA

The three essays collected in this book offer a succinct introduction to Agamben's recent work through an investigation of Foucault's notion of apparatus, a meditation on the intimate link of philosophy to friendship, and a reflection on the singular relation with one's own time that we call contemporariness.

Meridian: Crossing Aesthetics
$15.95 paper $50.00 cloth

Giorgio Agamben
A Critical Introduction

LELAND DE LA DURANTAYE

"Readers of Giorgio Agamben have long yearned for a guide to his work. This book is just such a guide: comprehensive, erudite, reliable, up-to-date, accessible, and properly critical."

—Wlad Godzic,
University of California, Santa Cruz

$24.95 paper $65.00 cloth

Enthusiasm
The Kantian Critique of History

JEAN-FRANÇOIS LYOTARD
Translated by
GEORGES VAN DEN ABBEELE

Enthusiasm studies what Kant calls a "strong" sense of the sublime, not as an aesthetic feeling but as a form of political judgment rendered not by the active participants in historical events but those who witness them from afar.

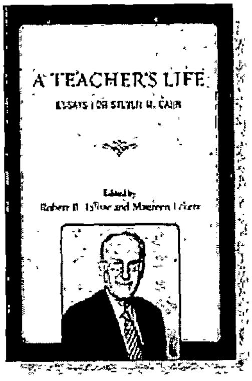
Cultural Memory in the Present
$18 95 paper $50.00 cloth

Stanford
University Press
800.621.2736
www.sup.org

Coming Soon from **Lexington Books**

A TEACHER'S LIFE
Essays for Steven M. Cahn

BY ROBERT TALISSE AND MAUREEN ECKERT



May 2009
978-0-7391-2288-4
$60.00 cloth

This is a collection of 13 essays honoring Steven Cahn, presented to him on the occasion of his 25th year as Professor of Philosophy at the City University of New York.

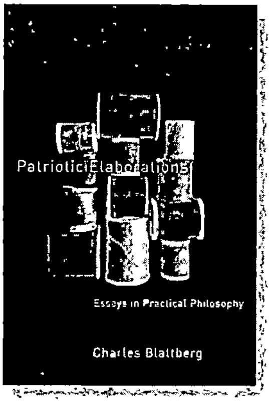
"This splendid collection develops philosophical and educational themes that have been the focus of Steven Cahn's attention throughout his remarkable career as a philosopher, educator, and academic leader." **—Karen Hanson, Indiana University**

Contributors: Norman Bowie, Steven M. Cahn, Randall Curren, Maureen Eckert, Alan Goldman, Tziporah Kasachkoff, Peter Markie, John O'Connor, David Rosenthal, Robert Simon, David Shatz, George Sher, Douglas Stalker, Robert B. Talisse, Christine Vitrano

An Imprint of the Rowman & Littlefield Publishing Group
To order visit **www.lexingtonbooks.com** or call **800-462-6420**

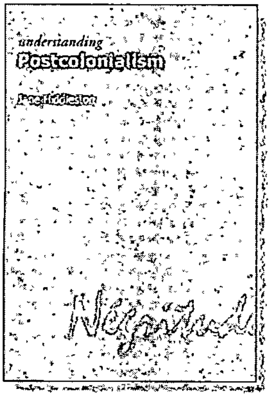
Truth and Truth-Making
Edited by E.J. Lowe and A. Rami
978-0-7735-3555-8 $27 95 paper | 978-0-7735-3554-1 $90.00 cloth

"This collection is very welcome and will be of great help to newcomers and to afficionados alike."
–Pascal Engel, University of Geneva



Patriotic Elaborations
Essays in Practical Philosophy
Charles Blattberg
978-0-7735-3491-9 $85.00 cloth

"A serious, thought-provoking, and interesting collection of insights that reflects a profound and intimate acquaintance with an impressive range of scholars, philosophers, artists, and authors. Blattberg is a creative thinker, often poetic."
–Dan Avnon, Hebrew University of Jerusalem



Understanding Postcolonialism
Jane Hiddleston
978-1-84465-161-0 $22.95 paper | 978-1-84465-160-3 $85.00 cloth

"A fresh and lucid presentation of major strands in postcolonial thought from Gandhi onward, including a clear presentation for the non-expert of the poststructuralist philosophical background necessary for understanding current debates."
–Lynda Lange, University of Toronto at Scarborough

"This is essential reading for all scholars and students in the postcolonial field, a concise and timely introduction to the philosophical underpinnings of anti-colonial and postcolonial criticism and a significant intervention in the field."
–Charles Forsdick, University of Liverpool

McGILL-QUEEN'S UNIVERSITY PRESS
www.mqup.ca

The MIT Press

THE MONSTROSITY OF CHRIST

Paradox or Dialectic?

Slavoj Žižek and John Milbank

edited by Creston Davis

A militant Marxist atheist and a "Radical Orthodox" Christian theologian square off on everything from the meaning of theology and Christ to the war machine of corporate mafia.

Short Circuits series, edited by Slavoj Žižek • 416 pp., $27.95 cloth

CONSCIOUSNESS REVISITED

Materialism without Phenomenal Concepts

Michael Tye

"This marvelously informed, powerfully argued book is Michael Tye's latest contribution to the task of finding a naturalistic understanding of consciousness."
— Frank Jackson, Princeton University

Representation and Mind series • 256 pp., 17 illus., $35 cloth

THE CRUCIBLE OF CONSCIOUSNESS

An Integrated Theory of Mind and Brain

Zoltan Torey

foreword by Daniel C. Dennett

"Just as poets often find that the constraints of rhyme and meter force them to discover strikingly apt expressions of their thoughts, it turns out that couching a computational theory of the mind in resolutely noncomputational terms pays dividends."
— Daniel C. Dennett, Tufts University and author of *Sweet Dreams*

272 pp., 2 illus., $34 paper

DISJUNCTIVISM

Contemporary Readings

edited by Alex Byrne and Heather Logue

"Disjunctivism is probably the single most important idea in philosophy of perception, epistemology, and theory of reference today. I don't know a better place to begin than with the cool, lucid overview provided by the editors, itself a masterly essay. An indispensable collection." — John Campbell, University of California, Berkeley

368 pp., $36 paper

BEYOND RED AND BLUE

How Twelve Political Philosophies Shape American Debates

Peter S. Wenz

"*Beyond Red and Blue* tackles tough issues from euthanasia to torture to global trade. Peter Wenz's conclusions may make you cheer or curse, but they are sure to make you think"
— Senator Dick Durbin, Illinois

376 pp., $27.95 cloth

CONCEPTUAL ANALYSIS AND PHILOSOPHICAL NATURALISM

edited by David Braddon-Mitchell
and Robert Nola

"In this excellent volume, [the editors] have brought together new essays by the most important contributors to the debate about the viability of the Canberra Plan, both defenders and skeptics. Anyone with a serious interest in metaphysics, a priori knowledge, the philosophy of mind, and meta-ethics will need to possess this book." — Tim Crane, University College London

A Bradford Book • 384 pp., $38 paper

THE TIME OF OUR LIVES

A Critical History of Temporality

David Couzens Hoy

"Drawing on years of immersion in the major traditions of continental philosophy, Hoy has produced a remarkable meditation on the implications of each for the perennially vexed question of temporality." — Martin Jay, University of California, Berkeley

328 pp., $34 cloth

PREDICATIVE MINDS

The Social Ontogeny of Propositional Thinking

Radu J. Bogdan

"Bogdan presents a highly original synthesis that draws on recent work in the philosophy of language, developmental psychology, evolutionary biology, and the philosophy of mind. This deserves to be an influential book"
— Robert M Gordon, University of Missouri, St. Louis

184 pp., 3 illus., $25 cloth

To order call 800-405-1619 • **http://mitpress.mit.edu**

The MIT Press

FUNCTIONS IN BIOLOGICAL AND ARTIFICIAL WORLDS

Comparative Philosophical Perspectives

edited by Ulrich Krohs and Peter Kroes

"The importance of [the book] extends beyond philosophy of biology and philosophy of technology into issues of general philosophical interest. Highly recommended" — Lynne Rudder Baker, University of Massachusetts Amherst

Vienna Series in Theoretical Biology • 312 pp., 12 illus., $50 cloth

THE ETHICS OF PROTOCELLS

Moral and Social Implications of Creating Life in the Laboratory

edited by Mark A Bedau and Emily C. Parke

Experts explore the potential benefits, risks, and moral aspects of protocell technology, which creates simple forms of life from nonliving material.

Basic Bioethics series • 392 pp., $28 paper

COGENT SCIENCE IN CONTEXT

The Science Wars, Argumentation Theory, and Habermas

William Rehg

"Rehg's riveting account of cogency will revolutionize the way critical theorists, logicians, and rhetoricians think about argumentation in general and, more particularly, about how we should assess the social institutionalization of collaborative research and debate in science proper for years to come." — David Ingram, Loyola University, Chicago

Studies in Contemporary German Social Thought
360 pp., 4 illus., $40 cloth

ZONE BOOKS

Now in Paper

THE CRADLE OF HUMANITY

Prehistoric Art and Culture

Georges Bataille

edited and introduced by Stuart Kendall
translated by Michelle and Stuart Kendall

A radically interdisciplinary inquiry into the origins of human consciousness, community, and potential.

Distributed for Zone Books • 224 pp., 15 illus., $19.95 paper

PHILOSOPHY OF LOVE

A Partial Summing-Up

Irving Singer

The author of the classic philosophical treatment of love reflects on the trajectory, over decades, of his thoughts on love and other topics.

The Irving Singer Library • 144 pp., $14.95 cloth

Now in Paper

THE PARALLAX VIEW

Slavoj Žižek

"No one demonstrates the continued philosophical vitality of Marxism better than Slavoj Žižek" — *Tikkun*

Short Circuits series, edited by Slavoj Žižek • 448 pp., $14.95 paper

Now in Paper

THE REALLY HARD PROBLEM

Meaning in a Material World

Owen Flanagan

"The book sparkles with thought and a likeable humour." — Steven Poole, *The Guardian*

304 pp., 1 illus., $15.95 paper

Now in Paper

NEUROPHILOSOPHY OF FREE WILL

From Libertarian Illusions to a Concept of Natural Autonomy

Henrik Walter

translated by Cynthia Klohr

"Walter uses contemporary findings in neurobiology and new understandings in the foundations of nature to explore 'free will.' The result is a fresh thesis well worth considering." — Kevin T. French, *Academia*

408 pp., 7 illus., $27 paper

Now in Paper

RATIONALITY AND LOGIC

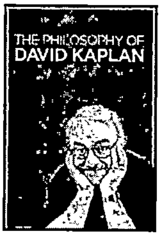
Robert Hanna

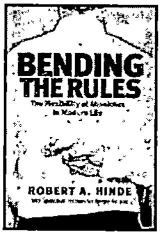
"This stimulating and wide-ranging book will be of interest to philosophers of logic and also to cognitive psychologists." — Jane Heal, University of Cambridge

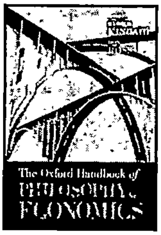
344 pp., $18 paper

To order call 800-405-1619 • **http://mitpress.mit.edu**

New from OXFORD







The Philosophy of David Kaplan
JOSEPH ALMOG and PAOLO LEONARDI
New, previously unpublished articles on the philosopher David Kaplan.
2009 | 336 pp | $99.00

Imposing Values
Liberalism and Regulation
N. SCOTT ARNOLD
An even-handed characterization of the differences between modern liberalism and classical liberalism.
2009 | 504 pp. | $74.00

Buddhist Philosophy
Essential Readings
Edited by WILLIAM EDELGLASS and JAY GARFIELD
An introduction to the vast and diverse traditions of Buddhist thought through a selection of their most significant texts.
2009 | 480 pp.
Hardback $99.00 | Paperback $24.95 |

The Oxford Handbook of Philosophy and Literature
Edited by RICHARD ELDRIDGE
23 newly commissioned essays by major philosophers and literary scholars that investigate literature as a form of attention to human life.
(Oxford Handbooks in Philosophy)
2009 | 536 pp. | $150.00

New in paperback!
Bending the Rules
The Twenty-First Century Morality
ROBERT A. HINDE
Robert Hinde argues that understanding the evolutionary origins of our morality and how we bend the rules can help guide us away from global catastrophe and towards a more ethical world.
2009 (Hardback 2007) | 256 pp.
Paperback $17.95 | Hardback $29.95

Our Stories
Essays on Life, Death, and Free Will
JOHN MARTIN FISCHER
John Martin Fischer defends the commonsense views that death can be bad and immortal life can be good, and argues that in acting freely, we transform our lives so that our stories matter.
2009 | 192 pp. | $65.00

My Way
Essays on Moral Responsibility
JOHN MARTIN FISCHER
"These essays, and Fischer's writings on free will and moral responsibility more generally, rank among the very best on this topic ever."
—Derk Pereboom, *Ethics*
2007 (Hardback 2006) | 304 pp.
Paperback $24.95 | Hardback $65.00

The Riddle of the World
A Reconsideration of Schopenhauer's Philosophy
BARBARA HANNAN
An introduction to the philosophy of Arthur Schopenhauer, written in a lively, personal style.
2009 | 176 pp.
Hardback $99.00 | Paperback $24.95

The Oxford Handbook of Philosophy of Economics
Edited by HAROLD KINCAID and DON ROSS
This volume is the first comprehensive, cohesive, and accessible reference source to the philosophy of economics, presenting important new scholarship by top scholars.
(Oxford Handbooks in Philosophy)
2009 | 688 pp. 13 line illustrations | $150.00

OXFORD
UNIVERSITY PRESS
To order, or for more information, please call 1-800-451-7556.
Visit our website at www.oup.com/us.

New from OXFORD

Sexual Solipsism

Philosophical Essays on Pornography and Objectification

RAE LANGTON

"These fine essays chisel away at patriarchal thought, especially its sometimes lazy defenses of free speech and shallow notions of autonomy. But they do so by wielding the sharpest tools in the kit of modern philosophy. Langton's conclusions challenge many liberals; her methods challenge many feminists. This is femininst philosophy at its best."
—Leslie Green, University of Oxford

2009 | 432 pp.
Hardback $100.00 | Paperback $40.00

Doing without Concepts

EDOUARD MACHERY

Edouard Machery argues that the dominant psychological theories of concept fail to provide a coherent framework to organize our extensive empirical knowledge about concepts.

2009 | 296 pp. | $65.00

The Oxford Handbook of Philosophy of Mind

Edited by BRIAN P. MCLAUGHLIN and ANSGAR BECKERMANN

An outstanding international team of contributors offer 45 specially written critical surveys of a wide range of topics relating to the mind.

(Oxford Handbooks in Philosophy)
2009 | 680 pp. | $150.00

The Moral Skeptic

ANITA M. SUPERSON

Aims to show the ways in which philosophers should broaden, narrow, and otherwise fine-tune the skeptic's position so that a defeat of skepticism is complete.

2009 | 264 pp.
Hardback $99.00 | Paperback $24.95

Toward a Humanist Justice

The Political Philosophy of Susan Moller Okin

Edited by DEBRA SATZ and ROB REICH

"This is a wonderful collection of essays by very distinguished authors on important aspects of justice illuminated by Susan Moller Okin's path-breaking work."—Amartya Sen, Harvard University

2009 | 272 pp. | $85.00

Beauty

ROGER SCRUTON

"Highly accessible and offers readers a unique and well-argued approach to the concept of beauty."
—Scott Duimstra, *Library Journal*

2009 | 176 pp. | 25 halftones | $19.95

Thoughts

Papers on Mind, Meaning, and Modality

STEPHEN YABLO

A collection of twelve essays by Stephen Yablo which together constitute a modern-day examination of Cartesian themes in the metaphysics of mind.

2009 | 336 pp.
Hardback $100.00 | Paperback $50.00

Visit www.oup.com/us to sign up for our monthly philosophy e-newsletter

OXFORD
UNIVERSITY PRESS

To order, or for more information, please call 1-800-451-7556.
Visit our website at www.oup.com/us.

CAMBRIDGE

New and Noteworthy Titles

After the Holocaust
The Book of Job, Primo Levi, and the Path to Affliction
C. Fred Alford
$80.00: Hb: 978-0-521-76632-6
$23.99: Pb 978-0-521-74706-6:
256 pp.

Kant's Moral and Legal Philosophy
Edited by Karl Ameriks and Otfried Hoeffe
The German Philosophical Tradition
$85.00. Hb: 978-0-521-89871-3:
350 pp

Rights, Race, and Recognition
Derrick Darby
$90.00: Hb: 978-0-521-51540-5
$32 99 Pb: 978-0-521-73319-9:
206 pp.

Kierkegaard
An Introduction
C. Stephen Evans
$90.00: Hb: 978-0-521-87703-9
$27 99: Pb: 978-0-521-70041-2:
222 pp.

Genetic Analysis
A History of Genetic Thinking
Raphael Falk
Cambridge Studies in Philosophy and Biology
$90.00: Hb: 978-0-521-88418-1:
344 pp.

Gadamer and the Legacy of German Idealism
Kristin Gjesdal
Modern European Philosophy
$90.00. Hb: 978-0-521-50964-0:
254 pp

Prices subject to change

What Philosophers Know
Case Studies in Recent Analytic Philosophy
Gary Gutting
$90.00: Hb. 978-0-521-85621-8
$29 99 Pb: 978-0-521-67222-1:
264 pp.

Nietzsche: Writings from the Early Notebooks
Edited by Raymond Geuss and Alexander Nehamas
Translated by Ladislaus Löb
Cambridge Texts in the History of Philosophy
$85 00: Hb: 978-0-521-85584-6
$22.99: Pb 978-0-521-67180-4:
328 pp.

Kierkegaard: Concluding Unscientific Postscript
Edited and translated by Alastair Hannay
Cambridge Texts in the History of Philosophy
$110.00· Hb: 978-0-521-88247-7
$39.99: Pb: 978-0-521-70910-1:
582 pp.

The Philosophy of Death
Steven Luper
$90 00· Hb: 978-0-521-88249-1
$28.99: Pb: 978-0-521-70912-5:
264 pp.

German Idealism and the Concept of Punishment
Jean-Christophe Merle
Modern European Philosophy
$90.00: Hb: 978-0-521-88684-0:
224 pp.

Plato's Myths
Edited by Catalin Partenie
$99 00: Hb. 978-0-521-88790-8·
272 pp.

An Introduction to Decision Theory
Martin Peterson
Cambridge Introductions to Philosophy
$85.00: Hb: 978-0-521-88837-0
$26.99. Pb: 978-0-521-71654-3:
328 pp.

Kant's *Idea for a Universal History with a Cosmopolitan Aim*
A Critical Guide
Edited by Amélie Oksenberg Rorty and James Schmidt
Cambridge Critical Guides
$99.00: Hb: 978-0-521-87463-2.
270 pp.

The Cambridge Companion to Darwin
Second Edition
Edited by Jonathan Hodge and Gregory Radick
Cambridge Companions to Philosophy
$90 00: Hb: 978-0-521-88475-4
$34.99 Pb: 978-0-521-71184-5·
562 pp.

The Cambridge Companion to Philo
Edited by Adam Kamesar
Cambridge Companions to Philosophy
$85.00: Hb: 978-0-521-86090-1
$29.99: Pb· 978-0-521-67802-5:
288 pp.

The Cambridge Companion to Boethius
Edited by John Marenbon
Cambridge Companions to Philosophy
$95 00: Hb: 978-0-521-87266-9
$32.99: Pb: 978-0-521-69425-4:
372 pp.

www.cambridge.org/us
CAMBRIDGE
UNIVERSITY PRESS
1584 · 2009
425 YEARS OF CAMBRIDGE PRINTING AND PUBLISHING

CENTRAL LIBR
23.9.09
UNIVERSITY
CALC
CORE CAMPUS

THE JOURNAL OF PHILOSOPHY

VOLUME CVI, NUMBER 6
JUNE 2009

page
309 *Internalism, Externalism, and the Architecture of Justification* Alvin I. Goldman
339 *Resisting the Force of Argument* Jonathan E. Adler

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Michael J. Mooney; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Barbara Gimbel, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

---

THE JOURNAL OF PHILOSOPHY

2009

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2009 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 6, JUNE 2009

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA

INTERNALISM, EXTERNALISM, AND THE ARCHITECTURE OF JUSTIFICATION*

One of the liveliest and most persistent controversies in contemporary epistemology is the internalism/externalism dispute. Although first introduced in connection with knowledge,[1] the terms 'internalism' and 'externalism' rapidly came to be associated primarily with competing theories of justification.[2] A justification theory is classified as internalist or externalist as a function of how it answers the question: "What kinds of states of affairs determine, or make a difference to, the justificational status of a belief (or other doxastic attitude)?" Are these states of affairs internalist or externalist in character?

---

* Earlier versions of this paper were given at the University of Edinburgh (as the "Nature of Knowledge" lecture), State University of New York at Albany, Texas A&M University, University of Wisconsin (Madison), and Heinrich-Heine University, Duesseldorf (Germany). I received helpful comments from many members of these audiences. I am particularly indebted to Holly Smith, Dennis Whitcomb, Allan Sidelle, and Jim Pryor for extensive discussion, advice, and debate on many aspects of the paper.

[1] See D.M. Armstrong, *Belief, Truth and Knowledge* (New York: Cambridge, 1973).

[2] See William P. Alston, "Internalism and Externalism in Epistemology," *Philosophical Topics*, xiv (1986): 179–221, and "An Internalist Externalism," *Synthese*, lxxiv (1988): 265–83; Michael Bergmann, *Justification without Awareness: A Defense of Epistemic Externalism* (New York: Oxford, 2006); Laurence BonJour, "Externalist Theories of Empirical Knowledge," in Peter French, Thomas Uehling, Jr., and Howard Wettstein, eds., *Midwest Studies in Philosophy*, Volume 5, *Studies in Epistemology* (Minneapolis: Minnesota UP, 1980), pp. 53–73; Roderick M. Chisholm, *Theory of Knowledge* (Englewood-Cliffs, NJ: Prentice-Hall, 1966, 1st edition; 1977, 2nd edition; 1989, 3rd edition); Juan Comesana, "We Are (Almost) All Externalists Now," in John Hawthorne, ed., *Philosophical Perspectives*, Volume 19, *Epistemology* (Malden, MA: Blackwell, 2005), pp. 59–76; Earl Conee and Richard Feldman, *Evidentialism: Essays in Epistemology* (New York: Oxford, 2004); Richard Fumerton, "The Internalism/Externalism Controversy," in *Philosophical Perspectives*, Volume 2, *Epistemology* (Atascadero, CA: Ridgeview, 1988), pp. 443–61, and *Metaepistemology and Skepticism* (Lanham, MD: Rowman and Littlefield, 1995); Carl Ginet, *Knowledge, Perception and Memory* (Boston: Reidel, 1975); Alvin I. Goldman, "The Internalist Conception of Justification," in *Midwest Studies in Philosophy*, Volume 5, *Studies in Epistemology*, pp. 27–52, and "Internalism Exposed," this JOURNAL, xcvi, 6 (June 1999): 271–93; Hilary Kornblith, "How Internal Can You Get?" *Synthese*, lxxiv (1988): 313–27; and Alvin Plantinga, *Warrant: The Current Debate* (New York: Oxford, 1993).

0022-362X/09/0606/309-38

Factors that (help to) fix justificational status are generally called *justifiers*, or *J-factors*. So the central question is whether justifiers, or J-factors, have an internalist or externalist character. Numerous types of factors, however, might influence justificational status. What if some are internalist and others externalist? Which side wins, in this case, internalism or externalism? Call this a question about the *terms of engagement*. One configuration of the terms of engagement is *existential*: externalism wins if there is at least one externalist type of J-factor. Internalism wins only if *all* J-factors are internalist. A second possible configuration is *majoritarian*. That side wins that has a majority of types of J-factors of the kind it promotes. The most common terms of engagement are the first, existential kind. This is because internalists have been quite explicit in claiming that all J-factors are internalist. For example, Roderick Chisholm writes:

> A consequence of our "internalistic" theory of knowledge is that, if one is subject to an epistemic requirement at any time, then this requirement is imposed by the *conscious state* in which one happens to find oneself at that time.[3]

I shall argue, however, that it is relatively easy to show that at least one type of J-factor is externalist. So the existential configuration makes things too easy for externalism. To level the playing field a bit, the majoritarian configuration should also be considered. Applying the majoritarian approach hinges on the admittedly knotty question of how justifiers should be typed and thereby counted. Although this challenge might produce some bickering, I welcome the second configuration along with the first. Perhaps it takes an excess of confidence for an externalist like me to offer the majoritarian configuration as one of two reasonable ways to set the terms of engagement. But it is instructive to explore the internalist/externalist controversy from a variety of perspectives.

A further complication in setting up the debate is that epistemologists commonly admit two conceptions of justifiedness: *doxastic* and *propositional*. Under the first conception, justifiedness and unjustifiedness are properties of doxastic state tokens (beliefs, suspensions of judgment, degrees of credence, and so on). Under the second conception, justifiedness and unjustifiedness are properties of ordered triples, consisting of a subject, a proposition, and a time. Even if person *S* does not believe proposition *P* at time *t*, nonetheless, the person might be justified *in* believing *P* at *t* (given his/her evidence). It is unclear whether the J-factors for doxastic justifiedness coincide with those for propositional justifiedness. Most of my formulations are intended to be largely neutral between the two, although that is difficult to achieve without repeatedly offering alternative verbal formulations. I will not inflict that on the reader.

---

[3] Chisholm, *Theory of Knowledge*, 3rd edition, pp. 59–60.

Let us turn to the conception of a justifier. William Alston defines a justifier as "anything that affects the justification of a belief, positively or negatively."[4] Read strictly, the definition applies only to beliefs. For most of the paper, however, I will adopt an expanded definition:

> (J) A justifier of any belief or other doxastic attitude is any property, condition, or state of affairs (and so on) that is positively or negatively relevant to the justificational status of that attitude.

How should we understand the term "relevance" here? It certainly should not be understood purely causally. Although some J-factors may be causally relevant to the belief in question, they are not causally relevant to the *J-status* of that belief. We may say of a J-factor that it "helps make" the belief's justificational status positive or negative, but the "making" in question is not the usual causal making. It is better to think of the relevant kind of relevance as *explanatory* relevance. *X* is a J-factor of a given belief's justificational status if and only if *X helps explain* why the belief's justificational status is what it is.

Next let us turn to the two leading conceptions of what it means for a J-factor to be internal as opposed to external. These conceptions are *accessibilism* and *mentalism*. There is considerable variation in how accessibilism is spelled out by different writers. The following characterization is intended to convey the core of the idea.

> *Accessibilist Internalism.* A factor or condition *J* is an internal justifier for *S*'s having doxastic attitude *D* vis-à-vis *p* (at *t*) if and only if
>
> (i) *J* obtains (at time *t*),
> (ii) *J* is positively or negatively relevant to the justificational status of *S*'s holding *D* vis-à-vis *p* (at *t*), and
> (iii) *J* is "directly" accessible to *S* at *t*—that is, *S* is capable of truly believing or knowing, at *t*, that *J* obtains, and is capable of knowing this "directly."

> *Mentalist Internalism.* A factor or condition *J* is an internal justifier for *S*'s having doxastic attitude *D* vis-à-vis *p* (at *t*) if and only if
>
> (i) *J* is a mental state, event, or condition of *S*, obtaining or occurring at time *t*, and
> (ii) *J* is positively or negatively relevant to the justificational status of *S*'s holding *D* vis-à-vis *p* (at time *t*).

---

[4] Alston, "Internalism and Externalism in Epistemology," reprinted in *Epistemic Justification* (Ithaca: Cornell, 1989), p. 192.

Accessibilism is culled from a variety of authors, including Chisholm, Carl Ginet, and Kent Bach.[5] Mentalism is primarily derived from Earl Conee and Richard Feldman.[6] For each interpretation of 'internal', there is a correlated sense of 'external' that simply rejects the distinctively internalist constraint (clauses (i) and (iii) of the accessibilist definition and clause (i) of the mentalist definition) on what it takes to be a J-factor. An externalist type of J-factor is one that is positively or negatively relevant to a belief's justificational status without having to satisfy the distinctively internalist restriction(s).

A slight emendation of the foregoing characterization of mentalist internalism might be in order. Timothy Williamson argues that knowledge is a mental state.[7] But knowledge, being a factive state, is not purely internalist. So if mentalists agreed that knowledge is a mental state and that knowledge is a potential justifier, they would have to conclude that not all justifiers are internal. This is something mentalists would certainly resist. The remedy is straightforward, however. In the definition of mentalist internalism, simply substitute the following clause for clause (i): "*J* is a *nonfactive* mental state, event, or condition of *S* (occurring at time *t*)." This is the interpretation of mentalist internalism I shall henceforth adopt—although I will not bother to insert the nonfactiveness qualification.[8]

---

[5] Chisholm writes: "the things we know are justified for us in the following sense: *we* can know what it is, on any occasion, that constitutes our grounds, or reason, or evidence for thinking that we know"—*Theory of Knowledge*, 2nd Edition, p. 252. Ginet writes: "Every one of every set of facts about *S*'s position that minimally suffices to make *S*, at a given time, justified in being confident that *p* must be *directly recognizable* to *S* at that time. By 'directly recognizable', I mean this: if a certain fact obtains, then it is directly recognizable to *S* if and only if, provided that *S* at that time has the concept of that sort of fact, *S* needs at that time only to reflect clear-headedly on the question of whether or not that fact obtains in order to know that is does obtain"—*Knowledge, Perception and Memory*, p. 34. Kent Bach writes: "Internalism ... treats justifiedness as a purely internal matter: if *p* is justified for *S*, then *S* must be aware (or at least be immediately capable of being aware) of what makes it justified and why"—"A Rationale for Reliabilism," *The Monist*, LXVII (1985): 246–63, p. 250.

[6] Conee and Feldman write that according to mentalist internalism, "internalism is the view that a person's beliefs are justified only by things that are internal to the person's mental life .... A mentalist theory may assert that justification is determined entirely by occurrent mental factors, or by dispositional ones as well. As long as the things that are said to contribute to justification are in the person's mind, the view qualifies as a version of mentalism." See Conee and Feldman, "Internalism Defended," in *Evidentialism: Essays in Epistemology* (New York: Oxford, 2004), pp. 53–80, on p. 55. Elsewhere Feldman and Conee indicate that only the mental states and conditions that obtain *at the time of belief* (or other doxastic attitude) are relevant to its justificational status: "The epistemic justification of anyone's doxastic attitude toward any proposition strongly supervenes on the evidence that the person has *at that time*." See Feldman and Conee, "Afterword to 'Evidentialism,'" in *Evidentialism: Essays in Epistemology*, pp. 101–07, on p. 101 (emphasis added). Note that this quotation occurs in a recent statement of their position.

[7] Williamson, *Knowledge and Its Limits* (New York: Oxford, 2000).

I. THE RIGHT-RULE ARCHITECTURE OF EPISTEMIC JUSTIFICATION

Let me now explain the phrase 'architecture of justification' that appears in the title. The architecture I have in mind is inspired by the uncontroversial fact that 'justification' is a normative term. Calling a belief justified might be compared to calling an act right, permitted, or perhaps even obligatory (in the circumstances). When we enter a normative domain, it is often appealing to think in terms of rules or principles that govern the domain in question. In the moral domain, people talk about moral rules or principles. Arguably, actions attain the status of being morally right or wrong by a combination of what the agent does, what the circumstances are, and what the governing moral rules prescribe, permit, or prohibit. Similarly, actions are legal or illegal because of what those actions consist in, what the circumstances are, and what the governing legal rules prescribe, permit, or prohibit. Finally, by analogy, perhaps belief tokens (and other doxastic-state tokens) are made epistemically right or wrong—in the sense of justified or unjustified—in virtue of the conditions, circumstances, or causes of their occurrence plus the governing epistemic rules. Numerous epistemologists propound epistemic rules and principles in precisely this spirit. Rules of epistemic rationality (for example, Bayesian rules) are more commonly displayed in the literature than rules of justifiedness, but there is plenty of precedent for justificational rules as well.[9]

In all such normative domains, the power to confer (objective) rightness or wrongness on a piece of "conduct" (behavioral or doxastic) is not possessed by any random rule. Such power is possessed only by *right*, or *correct*, rules. Anyone can propound random moral or legal rules, but the acts that fall under them will not *be* (objectively) right or wrong unless some *correct*, *authorized*, or *legitimate* rule implies that they are wrong. Why should things be different in epistemology? So I shall assume that only *right* epistemic rules make a difference to genuine justifiedness. This point should be equally acceptable to both internalism and externalism.

---

[8] There is also the question of whether an externalist approach to mental *content* renders mental states noninternal. I will not try to tackle this problem systematically because it is not *my* problem; it is a problem for internalism. However, note 126 below offers a few additional comments on it.

[9] Chisholm propounded principles of justifiedness (or epistemic statuses akin to justifiedness) in all three editions of his *Theory of Knowledge* (*op. cit.*). I have advanced a right-rule framework for thinking about justifiedness in two previous works: Goldman, *Epistemology and Cognition* (Cambridge: Harvard, 1986), and Goldman, "Epistemic Relativism and Peer Disagreement," in R. Feldman and F. Warfield, eds., *Disagreement* (New York: Oxford, in press).

Once this theoretical fact is acknowledged, two important questions belong on the table. First, which rules of justifiedness (J-rules) are correct or right? Second, what is the *criterion*, *standard*, *ground*, or *basis* for the rightness of J-rules? Illustrative answers to such questions will be discussed as we proceed. Staying at the abstract level for now, however, we can formulate the following *linkage principle*, which articulates what I have been calling a right-rule architecture. This linkage principle is also intended to be neutral as between internalism and externalism.

(L) *S* is justified in holding doxastic attitude *D* toward proposition *p* at time *t* if and only if there are some conditions *C* such that

(i) *S* is in conditions *C* at (or before) *t*, and
(ii) one or more right J-rules jointly permit a subject who is (or was) in *C* to form or retain attitude *D* toward *p* at *t*.[10]

Given principle (L), the justificational status of a subject's doxastic attitude toward a proposition is a matter of whether the attitude *conforms* to what right J-rules authorize in light of the subject's epistemic situation or activity. Conforming to rules is here contrasted with *following* rules, where rule following consists in conforming to rules as a result of *mentally representing* them and "verifying" that the adopted act or attitude conforms to them. I shall not consider a rule-following linkage principle at all. Some theorists might prefer a slightly different type of linkage principle than (L) (such as a rule-following principle), but I shall stick with (L).[11] The remainder of the paper examines prospects for epistemic internalism versus externalism in the context of the right-rule architecture articulated by (L).

II. J-RULES AND TYPES OF JUSTIFIERS

To relate the right-rule architecture to the internalism/externalism dispute, I proceed as follows. First, I examine the *kinds* of J-rules that hold initial promise of having the right form and general content, according to familiar epistemological theories. The antecedents of the rule schemata to be considered will specify types of conditions declared pertinent to the permissiveness of adopting certain doxastic attitudes toward certain propositions. If a rule schema is on target, material specified in the antecedent of one of its instances will be satisfiable by various types of *conditions, circumstances, states of affairs, facts,* or *processes.* In virtue of satisfying the antecedent of a right J-rule, these conditions, states of affairs, facts, and so on, will qualify as justifiers. Ultimately, however, they qualify as justifiers because they are *relevant* to the justificational status of various doxastic attitudes, as our definition of 'justifier' stipulates. However, not *all* justifiers qualify as such in virtue of satisfying antecedents of right J-rules. In addition to justifiers that satisfy antecedents of (right) J-rules, I shall later identify justifiers that include, among other things, right J-rules themselves, as well as the contents of the criterion, standard, or ground of rightness. So J-factors can qualify as such in a variety of ways. For now, however, I focus on the case of J-factors that qualify as such by satisfying antecedents of right rules. A central thesis of the paper is that many internalism-friendly rule schemas would need to be replaced by externalism-friendly rule schemas in order to be right.

---

[10] An alternative linkage principle might appeal to the cognitive *processes* a subject uses to arrive at the doxastic attitude in question. Right rules would authorize the use of selected cognitive processes. This kind of linkage principle is what I favored in *Epistemology and Cognition*. However, virtually all internalists would find linkage principle (L) more congenial than a cognitive-process linkage principle. Since I am trying to make the architecture as friendly as possible to internalism, so as not to prejudice my case against it, I shall proceed in terms of (L).

[11] Paul Boghossian argues that the rule-following conception of epistemic justification (or rational belief) is infected by a vicious regress. See his "Epistemic Rules," this JOURNAL, CV, 9 (September 2008): 472–500. This kind of vicious regress does not threaten (L), however. Some possible reservations about (L) emerge in my own paper, "Epistemic Relativism and Peer Disagreement." But I set those reservations aside for present purposes.

It might help the reader if I say a word about how I arrived at the present form of the argument. I originally planned to develop a hybrid position, one that would espouse a synthesis, or blend, of internalism and externalism. It would concede that internalism is right about those justifiers that satisfy the antecedents of right rules; they are indeed internal. But externalism is right about the contents of the *standard* of J-rule rightness. So internalism and externalism are each right about something important. On further reflection, however, I realized that the first thesis is incorrect. For the most part, right rules need antecedents that do not restrict the indicated conditions to internal factors. Thus, the position I adopt here is not very hybrid—in keeping with my past tendency to favor externalism over internalism.

I begin by distinguishing *inferential* and *noninferential* J-rules. Inferential rules license new beliefs (and other doxastic attitudes) based on logical, inductive, abductive, or probabilistic relations to pre-existing beliefs (and so on). Noninferential rules license new beliefs in virtue of the occurrence of nondoxastic states or events such as perceptual experiences and apparent memories. Of course, any such rule will specify conditions that permit the adoption or retention of doxastic attitudes in very general terms, so all of the illustrations below (and the ones that follow) are really just rule schemas. Particular rules would have to supply more concrete content to replace (many of) the schematic letters.

An inferential rule would take approximately the following form:

> (INF) If subject *S*, at time *t*, holds doxastic attitudes $D_1, D_2, \ldots, D_m$ toward propositions *q*, *r*, ..., *s*, respectively, and if proposition *p* bears relation *R* to the conjunction of *q*, *r*, ..., *s*, then *S* is permitted to adopt attitude D_n toward proposition *p* at *t*.

To specify a precise rule of this sort, '*R*' would have to be replaced with a suitable relation that can hold between propositions, for example, the relation of deductive, inductive, or abductive support, or some sort of probabilistic relation.[12] The above schema could accommodate many different theories and approaches to good inference, for example, inference to the best explanation or Bayesian inference.

A noninferential rule would take roughly the following form:

> (NONINF) If subject *S*, at time *t*, is in nondoxastic mental conditions *X*, *Y*, and *Z*, then *S* is permitted to adopt doxastic attitude *D* toward proposition *p* at *t*.

Of course, these are not the only types of rule schemas that might have correct instances. In fact, these particular schemas might not subsume any appropriate rules at all. I start with these rule schemas because they seem to capture a good bit of (at least informal) epistemological writing, and also to be quite congenial to internalism. Whether they are ultimately acceptable as schemas for right rules, and whether they genuinely support internalism, remains to be seen. In fact, I shall argue that even as they stand, these rule schemas are not entirely congenial to internalism (at least in the case of the inferential rule schema). Furthermore, each needs to be revised in ways inimical to internalism.

Any right rule that might exemplify one of these schemas could well need a *ceteris paribus* clause, or a "prima facie" qualification. This is not so easy to express within a right-rule framework. We would need the notion of "prima facie permission," which is not a very standard operator. In the interest of simplicity, however, I shall largely set aside problems involving defeaters and defeasibility that invite incorporation of "prima facie" qualifications. There is enough to worry about without this wrinkle.[13]

---

[12] The sample inferential rule schema I have chosen is more congenial to propositional justifiedness than to doxastic justifiedness. Doxastic justifiedness would require something to be said about the basing relation in addition to a confirmational support relation. However, propositional justifiedness tends to be the favored type of justifiedness among internalists, so my choice again follows the plan of being as accommodating as possible to internalism.

III. INFERENTIAL J-RULES: THE SUPPORT RELATION

Let us begin with inferential rules. Our initial sample of an inference-rule schema has a conjunctive antecedent. The first conjunct refers to doxastic states that a subject is in, and the second conjunct refers to some relation *R* of entailment, inductive support, and so on, that holds between the propositional contents of those doxastic states and the target proposition *p*. In considering whether any rules exemplifying this rule schema could both be right and meet internalist criteria, I begin with the second conjunct.

The fact that a certain logical, inductive, or probabilistic relation holds between specified propositions is a fact independent of any individual's mind.[14] The relation of entailment between the conjunction of 'All men are mortal' and 'Socrates is a man', on the one hand, and 'Socrates is mortal', on the other, is not a mental fact or condition. It is independent of any and all minds. Thus, factors or states of affairs that satisfy an antecedent of an inferential rule, under our initial proposal, will include states of affairs that are not *mental states* of subject *S*. These factors clearly will not qualify as internal according to mentalism. Would they qualify as internal according to accessibilism? Much depends on what accessibilists mean by "direct" access. Surely there are many relations of the intended type that are not directly accessible to most epistemic agents. Extremely complex entailment relations, holding between a huge set of complex premises and a conclusion, will not be "directly recognizable," or recognizable at all, for most epistemic agents. So even under the accessibilist notion of 'internal', many conditions or states of affairs that could satisfy the second conjunct of an antecedent of an inferential rule would not qualify as internalist J-factors.

Internalists sometimes respond to this problem by trying to "internalize" the ostensibly external facts that are candidates for justifiers. Conee and Feldman[15] present an example about a logic teaching assistant (TA) and a beginning logic student who are looking over a homework assignment together. One question in the assignment features a sentence that the TA and the beginning student both know to express a truth and asks whether certain other sentences are true as well. The student is clueless, whereas the TA can easily tell that some of the other sentences are logical consequences of the original one; so the TA is justified in believing that they are true as well. Conee and Feldman comment as follows:

---

[13] Our discussion of preservative memory in section v, however, will give some attention to defeasibility.

[14] See Goldman, "Internalism Exposed," and Comesana, "We Are (Almost) All Externalists Now."

[15] Conee and Feldman, "Internalism Defended."

> Again there is an internal difference between the two [the TA and the beginning student]. The difference is that the TA has justification for her beliefs to the effect that certain propositions validly follow from the original ones. She is expert enough to "see" that the conclusions follow without performing any computations.... [J]ust as in example 3, relevant internal differences make the difference (*ibid.*, p. 60).

The moral they draw from this and similar cases is the following:

> It is reasonable to generalize from these examples to the conclusion that every variety of change that brings about or enhances justification either *internalizes an external fact* or makes a purely internal difference. It appears that there is no need to appeal to anything extramental to explain any justificatory difference. These considerations argue for the general internalist thesis that these epistemic differences have an entirely mental origin (*ibid.*, p. 61, emphasis added).

What do Conee and Feldman mean by "internalizes an external fact"? A natural interpretation is this: S internalizes an external fact that p just in case S believes the (true) proposition p. Pursuing this hint, Conee and Feldman might ask us to rewrite our schema for inferential rules as follows, where the new version adds a belief operator to the second conjunct of the antecedent:

> (INF*) If subject S, at time t, holds doxastic attitudes D_1, D_2, ..., D_m toward propositions q, r, ..., s, respectively, and if S *believes that* proposition p bears relation R to the conjunction of q, r, ..., s, then S is permitted to adopt attitude D_n toward proposition p at t.

This rewriting of the rule schema eliminates reference to any external fact that posed a problem for internalism. Unfortunately, (INF*) is both *too strong* and *too weak.*

It is too strong because it is implausible to require people to have *beliefs* about logical, inductive, or probabilifying relationships as a precondition of having justified belief in conclusions that "rely" on such relationships. To get justified beliefs in those conclusions, it is sufficient to use psychological operations or computations that "realize" or "incorporate" corresponding principles. Ordinary cognizers rarely have a theoretical grip on these principles, or even have the concepts (for example, validity) in which the principles or relations could be expressed. Hence, unlike the logic TA described by Conee and

Feldman, they lack beliefs in the appropriate principles (for example, principles about validity). Nonetheless, their inferred beliefs can be justified. So (INF*) places too stringent a demand on inference-based justifiedness.

(INF*) is too weak in a different respect. By requiring *mere* belief in the holding of the relevant relation, rules that fall under (INF*) would allow a conclusion to be justifiedly inferred even though belief in the authorizing relationship is either *false* or entirely *unjustified.* Either of these scenarios is inappropriate. Conee and Feldman tacitly acknowledge the point about justifiedness. In their discussion of the TA example, they describe the TA (in part) as follows: "the TA has *justification* for her beliefs to the effect that certain propositions validly follow from the original one" (*ibid.*, p. 60, emphasis added). This sentence implies (or implicates) that *unjustifiedly* believing that certain propositions validly follow from the original one would not guarantee a justified belief in the conclusion ("such-and-such a sentence is true").

(INF*) is also too lax because it does not require an objective, "external" relation of validity or strong support to hold between the contents of the antecedently held beliefs and a newly inferred conclusion. Suppose someone thinks that affirming the consequent is a valid form of inference, and derives a belief in some proposition based on affirming the consequent. Clearly, his newly formed belief is not justified. But it would be justified according to (INF*). If there is no ("external") logical, inductive, or probabilistic *fact* of an appropriate sort, simply believing that there is one does not confer justifiedness on a resulting conclusion. Thus, the problem cannot be solved by formulating J-rules that "internalize" external facts, as Conee and Feldman express it. Right J-rules must feature objective, external (logical or confirmational) relations, a requirement that cuts against (strong) internalism of all stripes.[16]

In response, it might be argued that right inferential rules do not need to feature any conjunct of the second kind in their antecedents.[17] The right sort of formal relationship between antecedent beliefs (premises) and conclusion beliefs can be built into the antecedent and consequent of the J-rule schema itself. For example, there could be a J-rule schema of the following sort: "If *S* holds a pair of beliefs of the forms *Q* and (if *Q* then *P*), then *S* is permitted to adopt a belief in a proposition of the form *P*." This would implicitly capture the pertinent formal relationship and eliminate the need to specify any relation *R* in the antecedent.

---

[16] Subjective Bayesianism is surely a form of internalism, so does it not internalize the *R*-relation? No, I would respond. Subjective Bayesianism retains the "external" requirement that all rational or justified changes in degrees of belief must conform to objective truths of the probability calculus, especially Bayes's Theorem.

[17] This was suggested by Alan Sidelle, in a conference commentary on the paper at the University of Wisconsin, Madison, and something similar was suggested by Thomas Grundmann, at a workshop in Duesseldorf, Germany.

Suppose for the sake of argument that this is correct, and that the requisite content relationships can be adequately expressed in right inferential rules themselves, allowing us to dispense with the second conjunct in our original (or revised) rule schema. There is still another category of external facts that will play a justificatory role. This category of facts I shall call *instantiation facts.*

Melanie believes that Albert has just testified to *P*, that he has no motives for being disingenuous about *P*, and so forth. Does Melanie's cognitive situation permit her to infer *P*? Suppose there is a right inferential rule that authorizes one to believe a proposition if it is the best explanation of other things one already believes. And suppose that the truth of *P* is, as a matter of fact, the best explanation of Albert's testifying to *P* and other things Melanie believes about Albert.[18] Then Melanie is indeed justified in believing *P*. What facts or states of affairs are relevant to her being so justified? They include the antecedent beliefs from which Melanie infers *P*. Another such fact, however, is the fact that Melanie's antecedent beliefs and prospective conclusion belief (namely, *P*) jointly *instantiate* the right inferential rule we have posited. This instantiation fact is highly relevant to Melanie's J-status in believing *P* (where the J-status can be understood either in the doxastic or the propositional sense). If Melanie's cognitive situation did not instantiate this right inferential rule, believing *P* would not be justified for Melanie, because (we may suppose) no other right rule would license this belief. So the instantiation fact helps justify her (that is, helps make her justified) in believing *P*. This instantiation fact, however, is an external fact, not a mental fact. So its status as a justifier cannot be accommodated by mentalism. The same problem holds for accessibilism, because for most ordinary cognitive agents this kind of fact is not readily available for belief. Among other things, most such agents lack the concept of inference to the best explanation. The same general point can be made for other types of right inferential rules. Even if there is no conjunct in their antecedent that features an "R-relation," there is still another type of justifier that is externalist rather than internalist. These justifiers are not states of affairs that satisfy antecedents of right rules. But, as explained earlier, being a condition that satisfies the antecedent of a right rule is only one way, not the only way, in which a condition, state of affairs, or fact can qualify as a justifier.

---

[18] Peter Lipton defends an inference-to-the-best-explanation, or abductive, approach to the justification of testimonial belief, in "Alien Abduction: Inference to the Best Explanation and the Management of Testimony," *Episteme: A Journal of Social Epistemology*, IV (2007): 238–51. This article does not focus much on the *justificatory* properties of such an inference, but merely on the descriptive claim that this is how we make testimonial inferences. It is the justificatory angle that interests me, of course.

To avoid possible misunderstanding, I remind the reader that according to our linkage principle (L), a belief (or other doxastic attitude) is justified just in case it *actually conforms* to a right rule. (L) does not require that epistemic agents believe or know that their inferential behavior conforms to a right rule. Upon seeing this threat to their thesis, internalists might seek to internalize instantiation facts. Such an attempt would run into analogous problems identified earlier with attempts to internalize support relations.

IV. INFERENTIAL RULES: THE JUSTIFICATION CONDITION ON PRIOR BELIEFS

I turn now from the second conjunct of the antecedent of (INF) to the first conjunct. This conjunct has an important inadequacy. Epistemologists generally agree that prior beliefs (or other doxastic states) can only *transmit* justifiedness to new beliefs; they cannot *generate* such justifiedness. If initial beliefs lack justificational "juice," no such juice can be transmitted to a new belief. Thus, subjects starting with unjustified beliefs as premises cannot arrive at a new justified belief by *any* form of inference, no matter how impeccable. Hence, the first conjunct of the antecedent of (INF) must be revised to make it a model for genuinely correct J-rules. It must read roughly as follows:

(INF**) If subject S, at time t, *justifiedly* holds doxastic attitudes D_1, D_2, ..., D_m toward propositions q, r, ..., s, respectively, and if proposition p bears relation R to the conjunction of q, r, ..., s, then S is permitted to adopt attitude D_n toward proposition p at t.

This amendment poses a new problem for the thesis that all justifiers are internal. Having a doxastic attitude toward a proposition is uncontroversially an internal matter. But I want to question whether *justifiedly* having a doxastic attitude is equally an internal matter. If it is an external matter, this will be a big problem for internalism. I contend that all such J-factors are external.

On what grounds do I base the suggestion that justifiedly believing a proposition is an external rather than an internal matter? Does this not beg the central question between internalism and externalism?

Internalists like Conee and Feldman claim that justifiedness supervenes on mental states that occur at the time of belief.[19] Does this not imply that justifiedly believing a proposition is an internal matter? No, I reply. To say that something *supervenes* on states of kind *K* does not imply that it *is* a state of kind *K*. So even on the Conee-Feldman approach, justifiedly believing is not a mental state. However, if Conee and Feldman are right in holding that justifiedness supervenes on current mental states, does that not suffice for their purposes? If justifiedness supervenes on current mental states, then there will be a set of current mental states—call it *M*—on which the justifiedness of (token) attitudes $D_1, D_2, ..., D_m$ toward $q, r, ..., s$ (respectively) supervenes. If the justifiedness of these attitudes supervenes on *M*, does not the justifiedness of attitude D_n vis-à-vis *p* also supervene on *M*?

I shall not try to assess the truth of this last claim. Instead, I shall argue directly against the thesis that justifiedness supervenes on mental states obtaining at the time of belief (or states directly accessible at the time of belief). Specifically, I shall argue that the J-status of a belief held at time *t* partly depends, in the general case, on what transpired in the subject's cognitive history prior to *t*. Hence, some of the J-factors relevant to a given belief held at *t* will be states of affairs obtaining prior to *t*. The *historicity* of justifiedness is a major problem for internalism, because internalism maintains that all J-factors for a doxastic attitude held at time *t* are states of affairs that obtain at *t*.

V. PRESERVATIVE MEMORY AND THE HISTORICITY OF JUSTIFIEDNESS

The argument for the historicity thesis hinges crucially on the role of preservative memory in helping to determine a belief's J-status. To establish this point, we need an excursus into the epistemology of memory. One popular theory in the epistemology of memory is the "foundational" theory, which is very congenial to internalism. According to this view, conscious events of apparent recall provide foundational (prima facie) justification for believing the content of the recall. Just as a perceptual experience that *P* makes one prima facie justified in believing *P*, so an apparent memory experience that *Q* makes one prima facie justified in believing *Q*. Pollock and Cruz are among those who endorse this view.[20] On this theory, memory serves as a generative source of justifiedness, not merely as a mechanism that transmits justification (like inference). The foundational theory is congenial to internalism because a recall event at time *t* generates justifiedness for believing its content at the same time *t*. The J-factor (the recall event) is simultaneous with the (dated) J-status to which it is relevant, as internalism requires.

---

[19] Conee and Feldman, *op. cit.*

[20] John Pollock and Joseph Cruz, *Contemporary Theories of Knowledge* (Lanham, MD: Rowman and Littlefield, 1999, 2nd edition).

Michael Huemer identifies two problems with the foundational theory.[21] Suppose I adopt a belief in *P* in an unjustified fashion, for example, by wishful thinking. The next day I seem to remember that *P* but have no recollection of how I originally formed it. According to the foundational theory, assuming there are no defeaters, I am now justified in believing *P*. But that seems counterintuitive: Why should the mere passage of time—plus some forgetfulness on my part—suddenly make me justified in believing *P* although initially I was unjustified? Second, suppose I initially learn that *P* by means of a proof. Given the possibility of error, the justification is less than conclusive, but still quite adequate. After a few moments, however, I am now able to recall that *P*, while still entertaining the proof. So I will now have *two* justifications for *P*, one inferential and one foundational, thereby increasing my level of justifiedness vis-à-vis *P*. This, says Huemer, seems wrong. Why should the mere passage of time increase my justifiedness?

An alternative approach to memory's role in justification is the "preservation" theory. This theory does not much concern *conscious* memory, that is, conscious events of apparent recall. It mainly concerns how memory *storage* (an unconscious process) preserves justifiedness (or unjustifiedness) over time. It transmits justification rather than generating or creating it *de nuovo*. Preservative memory is what enables me to continue having justification for—and often knowledge of—the innumerable facts that I once learned but for which I no longer recall the evidence. For example, I now know that the sun is approximately 93 million miles from the earth. Assuming that knowledge entails justified belief (which virtually all internalists accept), I must be justified in believing this. But I no longer recall the source from which I learned it. Nonetheless I am justified because, when I first learned the fact, I knew who the source was (a reliable one, presumably), and hence was justified. What keeps me justified in believing this, even now, is the preservation of the earlier justifiedness via memory. Thus, the following J-rule is plausibly a right J-rule:

(PM) If *S* has a justified attitude *D* toward proposition *P* at *t*, and if *S* retains attitude *D* toward *P* until the later time *t′*, via memory, then, *ceteris paribus*, *S* is still permitted to have attitude *D* toward *P* at *t′*.

---

[21] Huemer, "The Problem of Memory Knowledge," *Pacific Philosophical Quarterly*, LXXX (1999): 346–57.

The *ceteris paribus* phrase is intended to accommodate the possibility that new evidence is acquired between t and t', thereby raising or lowering the appropriate level of credence for S vis-à-vis P. In light of (PM), a person's earlier J-status toward a proposition can clearly be relevant to its later J-status. Furthermore, as the *ceteris paribus* clause is intended to imply, any evidence S observes between t and t' inclusive can affect the justificational status of S's attitude toward P at t'. That is, it is *relevant* to the justificational status of the attitude toward P at t'. Hence, under the definition of J-factors, all such evidence observations qualify as *J-factors* vis-à-vis S's attitude toward P at t'. Thus, events or states of affairs occurring or obtaining prior to the time of justifiedness in question (here, t') can be genuine J-factors. (Remember: being a J-factor only requires being positively or negatively *relevant* to justificational status, not being decisively determinative of it.) This is what I mean by the "historicity" of justifiedness, and it is clearly implied by preservative memory.

A further example will be instructive. Ursula acquires by testimony an initially justified belief in a certain generalization, G. This occurs at t_1. She subsequently encounters a mountain of counterexamples to G, each of which is very obvious. The joint effect of these counterexamples, at t_5, is the total defeat of Ursula's original justification for G. She is now justified in *rejecting* G, not in believing it; hence she is unjustified in believing it. Nonetheless, Ursula continues to believe G. She either fails to consider how the observed counterexamples bear on G (an epistemically culpable omission), or she considers their bearing but fails to appreciate their falsification of G. Much later, at t_{20}, Ursula continues to maintain the same high confidence in G as she has held right along. On the first version of the story, she merely retains this belief in storage, along with her innumerable other beliefs. (She has no conscious apparent recall of its being the case that G.) This is a clear case in which Ursula's belief in G at t_{20} is unjustified. This assessment holds even if we also stipulate that at t_{20} Ursula has long since forgotten the defeating evidence she encountered at t_5. Her belief in G became unjustified then and it has retained this J-status ever since.

Ursula's being unjustified in believing G at t_{20} is incompatible with mentalism because Ursula has no mental states at t_{20} that constitute defeaters of her original justification for G. All such defeating mental states—her observations of the counterexamples—lie in the forgotten past. Similarly, it is incompatible with accessibilism, because at t_{20} those past defeating states are not directly accessible to her. So the unjustifiedness of Ursula's belief in G (at t_{20}) cannot be accommodated by either brand of internalism. Preservative memory is a problem for internalism because it shows that justifiedness at a given time does not supervene on states obtaining at that time.

Now the vast preponderance of things we know at a given time depend for their justifiedness on the rightness of a rule like (PM). This holds especially for items of knowledge stored in memory. Moreover, stored knowledge constitutes the great bulk of our knowledge, because there are narrow limits to the number of things held in consciousness at any given moment. The rest of what we know must be in states of storage. Although we cannot exclude the possibility that a stored belief can have positive J-status at a given moment without application of rule (PM), this possibility is problematic (see below). It is highly likely that almost all justified stored beliefs depend for their justifiedness on (PM). Thus, rejecting (PM) (and similar rules) would open the door to rampant skepticism about our common-sense knowledge. This would bestow a huge and unwarranted gift on the skeptic.

Consider the consequences of rejecting (PM) and its ilk. If (PM) is rejected, there will be an epistemic phenomenon that I shall call *continual clearing of the justificational slate* (CCJS). There will be no carryover of a belief's J-status from one moment to the next. Assuming for simplicity discrete moments in time, after each such moment the epistemic slate is wiped clean. A person may mnemonically retain a belief from one moment to the next, but that retention makes no contribution to the justifiedness or unjustifiedness of the succeeding moment's belief. Its J-status is determined by other factors entirely. If CCJS holds—which is implied by the rejection of (PM)—it will have catastrophic results for the number of beliefs people are justified in holding at any time.

Perhaps this is not obvious. Could not many stored beliefs be supported evidentially (that is, inferentially) by other stored beliefs held at the same time? Each would receive its justification inferentially rather than mnemonically. (Another possibility is evidential support from conscious beliefs held at the time, but, as previously indicated, conscious beliefs are vastly fewer in number and not usually content-congruent with the wide assortment of stored beliefs we learned long ago, so they would not help very much.) Assistance from other stored beliefs, however, runs into problems if (PM) is out of the picture. Other stored beliefs are helpful only if they themselves are justified. How many of them will be justified if they no longer inherit justificational juice via memory preservation? CCJS again rears its ugly head. If the inferential regress is to be stopped along familiar foundationalist lines, it presumably must stop with beliefs based on current perceptions. But beliefs based directly on current perceptions are unlikely to be of much help for most of our stored beliefs. No belief based on a current perception confers much if any justification on my belief that the sun is 93 million miles from the earth, that Versailles is in France, that my social security number is *XYZ*, and so forth. Thus, we cannot be rescued from serious skepticism if rule (PM) and similar rules are rejected.

However, the preservative theory of memorial justification (which inspires (PM)) has been criticized. Huemer criticizes it by invoking Russell's five-minute hypothesis embellished with a twin case (*ibid.*). Mike1 is Huemer himself who lives in a "normal" world and is justified in believing things about the past, for example, that he ate a bagel this morning. Mike2 is somebody God created just five minutes ago, endowed with exactly the same memories and environment as Mike1, except that in Mike2's case, the memories that purport to extend beyond the five-minute window are false. Intuitively, says Huemer, Mike2 should have the same memorially justified beliefs as Mike1 has, but the preservation theory disallows this. According to the preservation theory, memory is not an originating source of justification; it can only preserve justification initially conferred by some other means. There was no other originating source of justification in the case of Mike2. So the preservation theory must decree that Mike2's bagel belief is not justified.

Another case of interest is a variant of the original Ursula case (call it Ursula2) in which she has, at t_{20}, a conscious memory experience of *G*'s being true. She remembers nothing about how she originally formed her belief in *G* or her subsequent observation of counterexamples to *G*. She just seems to remember that *G*. Some people say that Ursula2 is justified in believing *G* at t_{20}; or at least that the apparent memory muddies the water, making it less clear that her t_{20} belief is *un*justified.

What shall we say, then, about the preservation theory and rule (PM)? A solution Huemer offers is a "dualistic" one, which combines elements of the foundational theory and the preservation theory. Dualism says that a belief is justified full stop if and only if one had an adequate justification for adopting it at some point, and thenceforward was justified in retaining it. How does this handle the Mike2 case? By adding the "posit" that coming to believe something by seeming to remember it is an epistemically justified way of coming to believe it. This will allow Mike2 to be justified in believing that he ate a bagel this morning (though his mode of justification—coming to believe it by seeming to remember it—will be quite different from that of Mike1). In Ursula2, dualism's verdict seems to run in the other direction. She does not justifiably adopt *G* (for the first time) by seeming to remember it at t_{20}, because she already believed it before t_{20}. But nor does she have an adequate justification for adopting it and thenceforward remain justified in retaining it; the defeaters at t_5 prevent this from being true.

Huemer's dualism has some definite virtues. It captures at least some of the attractions of both the foundational theory and the preservation theory. Internalists cannot accept it, of course, because the preservative aspect of dualism still implies the historicity of justifiedness. They will defend their rejection of it by insisting that dualism makes the wrong verdict in Ursula2. Some form of dualism must be right, however, something that balances the tension between preservationist intuitions and the skeptical consequences of CCJS, on one hand, and the pull of apparent-recall intuitions, on the other.[22] For our purposes, the bottom line is this. Any general theory of the epistemology of memory must assign a substantial—indeed, very prominent—role to preservative memory. The details are not crucial here. As long as preservative memory receives its due, so that something like rule (PM) is right, the historicity of justifiedness follows. That is to say, for some cases of an attitude's justifiedness at a time *t*, certain events or states of affairs *relevant* to that attitude's justifiedness at *t* occur prior to *t*. Under our definition of a J-factor, it follows that those prior events or states of affairs are J-factors of the attitude's J-status at *t*. This implies historicity.

How can internalists cope with this problem? One option is to abandon the claim that justification supervenes on *current* mental states (or that J-factors must be directly accessible from the subject's current position). Simply get rid of the temporal restrictions embedded in our original articulations of internalism. This would amount, however, to an abandonment of traditional internalism. Traditional internalism, the kind inspired by Descartes, aspires to show how ordinary beliefs can be justified from the starting-point of one's current mental states and nothing more. "Solipsism of the moment" is the epistemic situation that launches classical skepticism. Many epistemologists (including internalists themselves) identify internalism with a kind of epistemology that takes classical skeptical challenges seriously.[23] To abandon the currency constraint, therefore, is to abandon a large part of the impetus for internalism. Moreover, a currency requirement is part of the characterization of their approach offered by many leading internalists. It is clearly enunciated in the passage quoted earlier from Chisholm:

---

[22] Huemer puts more flesh on his dualism by suggesting what amounts to a "vectorial" theory of justification (not his term). One dimension would involve a degree of justification associated with the adoption of a belief, a second dimension would involve a degree of justification associated with its retention, and the overall level of justification would be a product of those quantities. This is an interesting proposal (which probably does not handle the Ursula case, however), but the details are not relevant to us here.

[23] See Michael Bergmann, "Externalism and Skepticism," *Philosophical Review*, CIX (2000): 159–94; BonJour, *The Structure of Empirical Knowledge*, p. 37; Fumerton, *Metaepistemology and Skepticism*, p. 180; Stroud, "Understanding Human Knowledge in General," in M. Clay and K. Lehrer, eds., *Knowledge and Skepticism* (Boulder: Westview, 1989), pp. 31–50.

> A consequence of our "internalistic" theory of knowledge is that, if one is subject to an epistemic requirement at any time, then this requirement is imposed by the *conscious state* in which one happens to find oneself *at that time*.[24]

To abandon the currency requirement—"at that time"—would be a major departure from traditional internalism.

VI. NONINFERENTIAL J-RULES

Where does the discussion to this point leave us? The discussion of inferential J-rules has yielded important consequences for three categories of J-factors. Two of these categories, but not the third, are types of J-factors that would satisfy antecedents of (plausible samples) of right rules. First, there are states of affairs consisting of the obtaining of support relations; these would be externalist J-factors. (If no reference to support relations appears in inferential rules, however, this case is of no consequence.) Second, there are states of affairs consisting of the subject's premise beliefs being justified, states of affairs that would help satisfy the first conjunct of right rules. By the historicity of justifiedness, however, the justifiedness of these premise beliefs will be an externalist state of affairs. Thus, these J-factors will also be externalist. Third, *instantiation* facts—which will also be J-factors—are externalist facts. (Instantiation J-factors are not peculiar to inferential J-rules.) So, at least two and possibly three types of J-factors associated with inferential justifiedness will be externalist.

It should not be concluded that *all* J-factors associated with inferential beliefs are externalist. Although a J-factor that consists in a belief's being *justified* is externalist, an arguably distinct J-factor that consists in the subject's *having* that belief is not externalist.[25] Nonetheless, internalism is losing more battles over types of J-factors than its proponents would hope. Three out of four categories of J-factors for inferential J-rules (or two out of three, if support relations are omitted) are externalist.

What happens when we turn to noninferential J-rules? This is friendly territory for internalism, one would think. Are not nondoxastic experiences, for example, perceptual experiences, the paradigm case of internal events? And are not such experiences prime examples of things that will figure in the antecedents of noninferential J-rules? Finally, since perceptual experiences do not have to *be* justified in order to *confer* justifiedness, J-factors for noninferential belief do not encounter the earlier problem of justifiedness appearing in the antecedent. In this section we shall see if the sailing for internalism here is as clear as these considerations might portend.[26]

---

[24] Chisholm, *Theory of Knowledge*, 3rd Edition, pp. 59–60 (second emphasis added).

[25] This will depend on how finely we individuate states of affairs.

A sample noninferential J-rule might have the following content: "If *S* has a visual appearance as of seeing an airplane overhead, *S* is permitted to believe the proposition that an airplane is currently overhead." This might be generalized into the following rule schema:

> (PERC) If *S* has a perceptual experience as of *P*'s being the case, then *S* is permitted to believe that *P*.

No doubt, some kind of defeasibility clause will be needed, but we set aside that complication here.

This rule schema gets one thing right. It is a general schema for noninferential J-rules that uses a systematic connection between properties of a perceptual experience and a propositional content that would figure in a permitted attitude (for example, belief). It does this by having matched propositional content in the antecedent and consequent. Unless something like this is done, it is hard to see how to formulate a perception-invoking rule with suitable generality. At the same time, this strategy gives rise to difficulties.

A first difficulty is the Sellarsian dilemma.[27] The phrase "perceptual experience as of *P*'s being the case" seems to refer to a perceptual state with propositional content. This suggests that it is really some sort of disguised doxastic state. If so, it is arguable that such a state cannot confer justification on a belief that *P* except by inference. If so, the rule belongs in the category of inferential rather than noninferential rules. Furthermore, it is doubtful that this sort of state can confer justification on a belief unless it itself is justified. What could make it justified? We would need an additional class of perception-based rules, and it is far from transparent what they would consist in.

---

[26] Of course, there are many contemporary approaches to mental states and contents, including perceptual states and contents, that would automatically make perceptual experiences "external." As far as perceptual contents are concerned, a Russellian approach to content would imply that a perceived object is part of the content of the state of perceiving it. Since perceptual objects are usually external, the contents of the perceptual state would also be external. There is still an additional step to be taken from the *contents* of the states being external to the *states themselves* being external. Many philosophers seem ready to take that additional step, perhaps on the assumption that a state's content is intrinsic to the state itself. In the epistemological literature, however, these two steps are not taken as commonly as they are in the philosophy of mind literature. It is either assumed that the contents are not Russellian (for example, they are Fregean), or that content-externalism does not imply state-externalism, or both. These matters often remain unclarified. It is simply assumed that mental states themselves are internal. Similarly, I shall assume that perceptual experiences do not qualify as external simply in virtue of their contents being external.

[27] Wilfrid Sellars, "Empiricism and the Philosophy of Mind," in *Science, Perception, and Reality* (New York: Humanities Press, 1963), pp. 127–96.

Suppose we try to circumvent the preceding difficulty by holding that a perceptual experience "as of *P*'s being the case" is a nondoxastic state that is the categorical basis of a disposition to believe that *P*. The state's content is not identical to *P*; rather, it is some sort of perceptual, nonconceptual content. Nonetheless, the state tends to cause in the subject a belief that *P*. Although this circumvents the preceding problems, it is unclear whether the rule schema is right. Is an epistemic subject always justified in believing *P* if she is in a perceptual state that inclines her to believe that *P*?[28]

I argue to the contrary. Suppose Reginald is in a visual state that inclines him to believe an identity-ascribing proposition to someone he sees in a crowd. It might be the proposition, "That's my long-lost love, Penelope, in the crowd" or the proposition, "That's the guy who has been pursuing me with apparent murderous intent." In each variant of the case, I imagine, Reginald has a visual experience that is partly caused by emotion. In the first variant it is an emotion of desire and longing. Reginald is so anxious to find Penelope that his visual experience is "as of Penelope being in the crowd;" although the visual cues leading to this visual experience are not at all indicative of its being Penelope. In the second variant, fear of the pursuer distorts Reginald's perception, so that he has a visual experience "as of the murderous assailant being in the crowd," despite the poverty of the visual stimulus. In both cases, I submit, Reginald is not justified in holding the indicated belief.[29] This cuts against the rightness of the proposed rule schema (under the proffered interpretation). Could it be amended so as to obtain a right-rule schema?

What is needed to obtain an acceptable rule or rule schema is a characterization of the perceptual cues or operations that properly license belief in such a proposition. What might be examples of such cues or operations? Which perceptual cues would provide suitable perceptual *evidence* for what is believed? Here is an example, taken from a simpler, well-studied domain, namely, depth perception. The visual system has multiple types of cues by which it estimates an object's distance from the perceiver. Two such cues are *binocular disparity* (the difference in images presented to the two eyes because of their different perspectives on the object) and *eye convergence* (the angle between the two eyes' lines of sight, that is, the extent to which the two eyes are turned toward each other to fixate an object, especially a nearby object). These conditions provide evidence to the visual system in its computation of an object's distance. Perhaps a subject whose visual system undergoes the right sorts of binocular disparity cues or eye convergence cues is thereby justified in believing that the observed stimulus is a certain distance from him. Borrowing pertinent details from vision science, we might be able to construct a J-rule that fits the circumstances in which subjects make accurate distance judgments based on binocular disparity or eye convergence.[30]

---

[28] Of course, initial justifiedness could be defeated by concomitant beliefs. But that is not the kind of problem I have in mind here. I am not worried about mere defeasibility, but that even prima facie justifiedness is not guaranteed.

[29] Not long after incorporating these examples into this paper, I heard Susanna Siegel present a paper with similar examples, many of which she also considered to involve unjustified belief. See Siegel, "Cognitive Penetrability and Perceptual Justification" (unpublished, Harvard University).

Rules of this sort, however, would not be congenial to internalism—neither to accessibilism nor to mentalism. The problem for accessibilism is that the cues associated with binocular disparity and eye convergence are not conscious. Presumably they are not directly accessible to the subject. What about mentalism? It is not clear whether this is an equally severe problem for mentalism because it is not clear what qualifies as "mental." It is highly plausible, however, that being the mental state of a person requires more than being a representational state of some cognitive *module* in that person's brain. In most cases concerning cues or operations of the visual system that enable people to form accurate visual beliefs, these cues are only states or operations of one or more modules of the visual system, not ascribable to the person, and hence not mental states of the person.[31] If this is correct, it seems difficult if not impossible to construct any plausible perception-based J-rules with antecedents that meet internalist strictures.

The upshot is that identifying a noninferential J-rule schema that meets internalist desiderata is not a slam-dunk. We have not definitively shown that such a J-rule cannot be had. But the prospects are murkier than is commonly thought. Thus far, then, we have identified several categories of J-factors that are clearly externalist, only one category that is clearly internalist, and another (perceptual experiences) where the classification is not wholly resolved. Under the "existential" terms of engagement, in which externalism wins if there is even *one* externalist category of J-factors, externalism clearly wins. If we apply the majoritarian terms of engagement, under which externalism wins if there are more externalist than internalist categories of J-factors, externalism still wins at this stage. We are not finished, however. There are still additional categories of J-factors for which the analysis tilts toward externalism.

---

[30] Thanks to Jack Lyons for suggesting these examples. His *Perception and Basic Beliefs* (New York: Oxford, 2009) provides an illuminating and subtle analysis of closely related issues.

[31] For further discussion of this point, see Lyons (*loc. cit.*, chapter 3).

VII. OTHER EXTERNALIST J-FACTORS: RIGHT J-RULES AND THE STANDARD OF J-RULE RIGHTNESS

I shall now argue that under the right-rule architecture and our definition of justifiers, right J-rules should qualify as justifiers. The argument is fairly straightforward. If a particular belief is justified, and it is justified in virtue of instantiating a certain right J-rule, then the content and rightness of that J-rule are positively relevant to its justifiedness. Given our definition of justifiers, this implies that the J-rule is a justifier of the belief's justifiedness. The argument can be amplified a bit by recalling that the species of relevance in question is explanatory relevance. It is explanatorily relevant to the belief's being justified that it instantiates the rule in question and that this rule is right. The fact that the rule has a certain content and is right helps explain *why*, or how it transpires, that the belief is justified.

Some people might respond as follows: "It is a mistake to regard either J-rules themselves or their rightness as J-factors. What is relevant to a belief's justifiedness are states of affairs that satisfy the operative rule's antecedents and the associated instantiation fact. But it is a *category mistake* to add the rules themselves, or their rightness, to the list of a belief's J-factors."[32]

In reflecting on this matter, it may help to consider an analogue from another normative domain. When we ask whether an accused is guilty or innocent of a specified legal offense, what kinds of factors would (help) make it the case that he is guilty or innocent? Obviously, many material facts—what deeds were done, in what circumstances, in what state of mind, and so on—are relevant. In addition to these material facts, however, another group of relevant factors are legal facts. What is the content of the law concerning the offense in question? Is it a valid law (under the governing federal or state constitution)? What are the legal precedents? The material facts alone do not make it the case that the person is guilty or innocent. The material facts alone do not explain *why* the person is guilty (if he is). The contents of the law are equally relevant. In other words, legal rules are

---

[32] Holly Smith has defended this position in conversation.

(partial) "determiners" of guilt and innocence as much as material facts are (which is not to deny the differences between them).

Given that right J-rules are justifiers, is their being right an internalist or externalist matter? That depends on what makes a right J-rule right. J-rule rightness is something conferred by the criterion, or ground, of rightness. If a J-rule meets the criterion, or ground, it is right; if it does not meet the criterion, it is not right. Is meeting such a criterion an internalist matter? Let us examine this first under the mentalist approach and then under the accessibilist approach.

Is meeting a rightness criterion a matter of an individual subject's mental states? Presumably not. For any selected criterion, whatever its content, meeting or failing to meet the criterion is an objective matter of fact, not a matter of somebody's belief or other mental state. Is meeting a rightness criterion directly accessible to any pertinent subject? Direct accessibility consists of being able to determine correctly, though directly, that the state of affairs in question obtains. Is the rightness of a J-rule determinable correctly and directly? The question should be clarified: determinable by whom? I assume that the answer is: by the epistemic subject, a person on the street. The person on the street cannot decide correctly which J-rules are right. Ordinary people have never considered such specialized issues (neither have many philosophers). The person on the street never dreams of these questions, much less their answers. He or she would have no clue as to what a standard or criterion of J-rule rightness is; nor would he or she be in a position to determine which J-rules meet such a standard. Hence, on neither approach to internality does the rightness of a J-rule qualify as an internal state of affairs. In this category, like most of those examined earlier, J-factors turn out to be external rather than internal.

Finally, what about the criterion, standard, or ground of J-rule rightness? Is this criterion itself a justifier of beliefs and other doxastic attitudes? Again, the answer seems to be affirmative. After all, the justificational status of each token attitude is a function, positively or negatively, of a specific right J-rule, or the total set of right J-rules. But the membership of the set of right J-rules is a function of the standard, ground, or criterion of J-rule rightness. So, by transitivity, the justificational status of each token attitude depends on the standard, ground, or criterion of J-rule rightness. Hence, this standard falls in the category of a justifier.

Is this justifier internalist or externalist? That is, does the *content* of the rightness-standard invoke internalist or externalist properties or states of affairs? An obvious externalist candidate for the rightness standard would invoke *truth-conduciveness*—for example, conduciveness to a high truth-ratio (or veritistic value[33]) among the beliefs that would result from conforming to a rule. The standard might say, for example, that a J-rule *V* is right if and only if, among the J-rules that compete for a certain category of J-rules, conformity with *V* would maximize the truth-ratio (or veritistic value) of the doxastic attitudes that would be generated if epistemic subjects conformed their doxastic "conduct" to *V* (as compared with the truth-ratio that would be generated by conforming their doxastic conduct to alternative rules).

Up to this point I am not trying to argue that the (correct) standard of J-rule rightness *is* externalist. I am just trying to clarify the issue and sustain the thesis that the standard itself (that is, its content) is among the justifiers of token beliefs. The content of the standard is one of the things that (directly or indirectly) help to fix the justificational status of doxastic attitudes.

No doubt other candidates for the standard of J-rule rightness would feature internalist contents. However, it might take more than a little delicacy to formulate some plausible ones. Some internalists—namely, mentalists—would probably resist the idea that a rightness criterion is needed. Conee and Feldman defend internalism primarily through the vehicle of a supervenience principle: their principle "*S*."[34] They do not explicitly claim that this supervenience principle provides a *full* theory of justifiedness, but they strongly imply this. It is important to indicate, therefore, that this is incorrect. A supervenience principle like "*S*" does not provide everything a theorist wants to know about justifiedness.

To explain this important point, consider an analogy. Suppose a moral consequentialist holds that the moral rightness of actions supervenes on their consequences. Is this a theoretically complete story of moral rightness? Hardly. It leaves entirely open *which* (types of) consequences confer rightness. Is rightness fixed by the pleasure that actions cause, or by different types of consequences (for example, the surprise they occasion)? Even being told that rightness supervenes on hedonic consequences leaves crucial questions unanswered: Does *more* pleasure or *less* pleasure conduce to rightness? The answer might seem obvious. But supervenience per se only guarantees that if two items share the same base properties, they also share the same supervening properties. It does not necessarily tell us precisely *which* properties supervene on *which* base properties.

---

[33] What is meant by "veritistic value" is explained in my *Knowledge in a Social World* (New York: Oxford, 1999). See note 35 below.

[34] Conee and Feldman's supervenience principle *S* reads as follows: "The justificatory status of a person's doxastic attitudes strongly supervenes on the person's occurrent and dispositional mental states, events, and conditions" ("Internalism Defended," p. 56). This principle does not explicitly restrict the supervenience base to states occurring at the same time as the doxastic attitudes in question. Another recent statement of theirs, however, quoted above in note 6, indicates that they do intend this restriction.

It is similarly under-informative to be told that epistemic justifiedness supervenes on a subject's mental states. Even if this were true, it would not tell us everything relevant. We also need to be told which combinations of mental states give rise to *positive* justifiedness and which combinations give rise to *negative* justifiedness. Knowing that mental states $M_1, \ldots, M_k$ jointly fix the J-status of an attitude toward q would not indicate whether the resulting J-status is positive or negative. This is why right J-rules are needed, and, at a higher level, why a criterion of J-rule rightness is theoretically desirable. Specifying a supervenience base (for example, the class of mental states) leaves a great deal indeterminate. This indeterminacy is what a truth-conduciveness criterion, or other such criterion, needs to resolve. I do not mean to imply that no internalist rightness criterion could possibly be up to the challenge. But it must be shown that an internalist rightness criterion *is* up to the challenge. One cannot dismiss the challenge on the ground that no rightness criterion is needed.

I next sketch an argument in support of an externalist, truth-conduciveness approach as the right approach to the job. The argument is simply that such a criterion would provide a satisfying explanation of why epistemologists find certain candidate J-rules right (or roughly right) and others wrong. I shall illustrate the point first with respect to noninferential J-rules and then with respect to inferential J-rules.

Why do so many examples of noninferential J-principles center on perceptual experience, especially the epistemic subject is in "good" perceptual circumstances? Because these are cases in which beliefs formed in accordance with these experiences are usually true. Or because these are cases in which epistemologists assume, and assume that others assume, that believing in accordance with the recommended principles leads predominantly to truths. What I say is intended to apply not only to reliabilists, who explicitly invoke a truth-conduciveness story. I also speak of epistemologists who offer an entirely different theory, but still adduce similar principles. I claim that the underlying appeal of these J-principles is a tacit recognition that they are truth-conducive, even when this is not the *official* doctrine being endorsed.

A similar example can be drawn from inferential J-rules. Which R-relations do epistemologists regularly choose to incorporate into inferential rules? Relations such as deducibility, strong inductive support, and inference-to-the-best-explanation. We do not choose epistemic rules in which the support relation is weak probabilistic support, or inference-to-the-*worst*-explanation! Why not? Because these rules would yield a much lower rate of true-belief formation than the preferred principles. Or so it is assumed. What apparently underpins the choice of congenial J-rules is their tendency to produce a fairly high ratio of true beliefs. So that is the appropriate choice of a standard of J-rule rightness. Even if some of these rules are criticizable for one reason or another, the best unifying explanation of their appeal to epistemologists is that epistemologists expect good epistemic practice to be the kind of practice that is comparatively truth-conducive.[35]

If all this is correct, the J-factor at the level of the rightness criterion is an externalist one. This adds yet another type of externalist J-factor to our list, covering many categories of justification determiners. So externalism appears to win the dispute with internalism even when the playing field is made more congenial to internalism than it normally is.

We have not examined any possible rightness criteria from the internalist camp. One such candidate comes from Richard Foley.[36] According to Foley a belief is rational if and only if it conforms to the believer's most deeply held principles about promoting the epistemic goal of now believing the truth. Translating this into a proposal for justification rather than rationality, this suggests that a J-principle is right just in case it is one of the believer's most deeply held principles about getting the truth. Being deeply held certainly looks like an internalist matter. It takes little reflection, however, to notice that Foley's criterion is highly relativistic. A principle deeply held by one epistemic subject is not necessarily deeply held by another. This makes Foley's approach ill suited to the objectivist, nonrelativistic spirit of our entire framework. As explained at the outset, our right-rule architecture links justifiedness to conformity with *right* rules, and it looks as if Foley's approach gives no sense to person-independent rightness of rules. A given rule may be right-for-Smith but not right-for-Jones, with no sense made of rightness *simpliciter.* It is not clear, then, that one can treat Foley's criterion as a *rightness* criterion. Instead it is a *relative-rightness* criterion, a different kind of beast altogether.

---

[35] The explanation in the text cites only the propensity of the exhibited rules to produce true *categorical* beliefs. This cannot be the entire story, because many prominent epistemologies feature *graded* beliefs as appropriate outputs of good J-rules. I have sought to capture the value of graded beliefs in a truth-conduciveness framework by introducing the broader notion of *veritistic value*, in which even graded degrees of belief assigned to truths are good epistemic consequences (as compared with lower degrees of belief assigned to the same truths). For a systematic development of this approach, see my *Knowledge in a Social World*.

[36] Foley, *A Theory of Epistemic Rationality* (Cambridge: Harvard, 1987).

VIII. CAN INTERNALISM BE SAVED BY A NEW DEFINITION OF JUSTIFIERS?

After reviewing my case for externalism, readers might suspect that I have been unfair to internalism from the outset in my characterization of justifiers. I adopted Alston's characterization of justifiers as any states of affairs that are positively or negatively relevant to justification. This is too broad, it might be objected. In fact, Alston himself gives a rather different characterization of justifiers in another paper. In "An Internalist Externalism," Alston introduces the notion of a "ground" as a state on which a belief is "based" (*op. cit.*). Prime candidates for grounds are the subject's perceptual experiences and other beliefs. Suppose now that we restrict "justifiers" to *grounds* in Alston's sense. Would not this substantially change the terms of the debate, and would not internalism win the debate so defined? If the outcome of the debate is so easily changed, should not our original definition be deemed arbitrary and prejudicial?

There are several reasons not to choose grounds as the linchpin of justifiedness. First, a person can have a justified belief without having any ground at all for the belief. Someone can justifiably believe a certain factoid for which he once had evidence he has since forgotten entirely. Nothing qualifies as his (current) grounds for the proposition, but he still believes it justifiedly. True, some writers say that memory preserves his original justification, that is, his original grounds. Therefore, he must still "have" them. There is no pertinent sense of "possession," however, in which a person who has forgotten his original grounds still possesses them. This just shows that *being* justified cannot be equated with *having* a justification in the sense of having a ground. Notice, moreover, that if having had grounds in the past sufficed for justifiedness, this would provide no succor for internalism, because internalism denies epistemic force to the past.

Another reason not to privilege "grounds" as the only factor to consider in analyzing the contours of justifiedness is that, as we have seen, grounds are just one among many types of factors relevant to a belief's justifiedness. In addition to grounds, there are (1) support relations, (2) instantiation facts, (3) historical events that bear on premise-belief justifiedness, (4) right rules, and (5) the ground or standard of rightness. In addition to these, there is one we have not much discussed: cognitive processes that constitute or underlie the "basing" relation. The basing relation is critical for doxastic justifiedness in particular. If you justifiably believe both *Q* and (if *Q* then *P*), you have the "makings"

CENTRAL LIBRARY CALCUTTA

(grounds) for a justified belief in *P*. But even if you go ahead and believe *P* in these circumstances, it does not follow that your belief in *P* is doxastically justified. It would not be justified if you ignored *Q* and (if *Q* then *P*) and instead came to believe *P* by fallaciously inferring it from your prior beliefs in *R* and (if *P* then *R*). So, current grounds are not the be-all and end-all of justifiedness, and the dispute between internalism and externalism should not be equated with the dispute over the current-grounds factor in justifiedness.

IX. CONCLUSION

Two possible terms of engagement were delineated for addressing the internalism/externalism dispute: the existential and the majoritarian configurations. I have argued that there are numerous categories of J-factors that qualify as externalist, many more than J-factors that qualify as internalist. If at least one of those arguments is correct, externalism wins the dispute as judged by the existential terms of engagement. Even by the more stringent, majoritarian term of engagement, externalism comes out on the winning side of the ledger.

ALVIN I. GOLDMAN
Rutgers: The State University of New Jersey

RESISTING THE FORCE OF ARGUMENT*

Why is there persuasive force, even when not yielding conviction, in rhetorical questions of the form "How else, but because of *R*, do you explain that *p*?" The question presupposes that *p* calls for explanation, and that *R* is a (purported) explanation of it. A similar persuasive force is found for arguments because the "how else" challenge is standardly part of an argument to accept a conclusion (for its explanatory promise). In responding to an arguer's presentation, you should either refute the argument or else accept its conclusion (as true), since you grant the argument's cogency which, for brevity, let us understand as covering soundness of relevant, nonquestion-begging premises and the appropriate formal relation between premises and conclusion, dependent on the nature of the argument. I will refer to this putative norm as *refute-or-accept*, and concentrate on its role in argument, mostly setting aside the explanatory—"How else?"—analogue. (In my relaxed use of 'argument' it refers either to a product [the argument], typically as presented, or to the practice of engaging the argument in dialogue. Where misunderstanding lurks, 'argumentation' is for the latter.)

We seem to endorse refute-or-accept. Should we? Daniel Dennett reports:

> Surely the following has happened to you, it has happened to me many times: somebody corners me and proceeds to present me with an argument of great persuasiveness, of *irresistible* logic, step by step by step. I can think of nothing to say against any of the steps. I get to the conclusion, *but I don't believe it!* This can be a social problem. It is worse than unsatisfying to say: 'Sorry, I don't believe it, but I can't tell you why. I don't know.'[1]

---

*For their comments, my thanks to Bradley Armour-Garb, Catherine Elgin, Elizabeth Fricker, Arnon Kernen, Matthew Moore, David Owens, Georges Rey, Michael Rescorla, Lance Rips, Michael Stocker, Rivka Weinberg, the editors of this JOURNAL, and especially Matthew Kotzen. This paper received helpful comments from audience members who attended readings at the University of London conference "What Is Testimony?" (2005), at the American Philosophical Association Eastern Division meetings (2005), and at the University of California, Santa Barbara (2007). The commentators at the APA presentation, Simon Evnine and Walter Sinnott-Armstrong, provided many valuable criticisms and suggestions.

[1] Dennett, "How to Change Your Mind," in his *Brainstorms* (Montgomery, VT: Bradford, 1978), pp. 300–09, p. 308. Dennett introduces this reflection for very different purposes.

0022-362X/09/0606/339–64

I will construe Dennett's tale as standing for a common dilemma. Dennett's discomfort with his lack of belief is due to his endorsing something like refute-or-accept and his recognition that he is in violation of it. Since belief is the source of the discomfort, I call this dilemma the "resistance to argument by belief" dilemma or the *RAB-dilemma.*

For an example, take a widely held, unremarkable political belief like that in wealthy, democratic nations, taxes should be levied for public works and to help the needy. Many of those who hold this belief would be left, I suppose, without a serious response (even from their own point of view) against even a mildly sophisticated libertarian argument. Nevertheless, they resist not only persuasion, but suspension of judgment.

When I refer to the libertarian case, you can substitute any realistic and easily imagined instance of a RAB-dilemma. Ordinary citizens regularly read, or listen to, or are confronted by, serious arguments that draw conclusions contrary to their own well thought-out opinion. They are not only unable to refute them, but, like Dennett, they find them "step by step" persuasive. Yet, they remain unconvinced. If the resistance is reason-based, is it dogmatic or can it sometimes be justified? If the latter, what account justifies the resistance, while allowing for it to be intellectually discomforting?

The dilemma has been missed in part because the basis for the resistance is assimilated to *reductios.* Remarkably, as I observe below (section VI), the assimilation actually depends on something like refute-or-accept.

However, I am sure that you already dismiss that putative norm because of what I refer to as the *Lockean objection:*

> I grant that I cannot refute your argument [explain that *p* otherwise]. But that is no reason to accept yours. That I cannot refute your argument [explain that *p* otherwise] may just be a limit or fault in me. It has nothing to do with whether yours succeeds.

Effectively, this is Locke's response to the form of argument he identified as the *ad ignorantium:* "It proves not another man to be in the right way, nor that I ought to take the same with him, because I know not a better."[2]

I. ORGANIZATION

Sections II–VI defend, while modifying, both horns of the RAB-dilemma. One horn is refute-or-accept; the other is the conditions that entitle Dennett to resist believing. In the next section (II), I clarify refute-or-accept's claims and support, but also stress its strong demands. It places burdens on the option of suspension of judgment, treats acceptance as issuing in belief, and imposes restrictions on refutation. The following section (III), nevertheless, supports refute-or-accept further by drawing comparisons with testimony. Testimony is an excellent (reliable) source of warranted beliefs, governed by a parallel norm. Yet, testimony is subject to a variant of the Lockean objection. Still, as brought out in sections IV and V, our vulnerability to false persuasion is far greater for argument than for testimony. Section VI argues that resistance in RAB-dilemmas does not amount to reductios, so that this is not a viable way to reduce the vulnerability.

---

[2] Locke, *An Essay concerning Human Understanding,* P.H. Nidditch, ed. (New York: Oxford, 1975), Book IV, Chapter XVII, section 20. Arguments from ignorance often take the form of burden of proof reasoning, which is similar to refute-or-accept with a difference that I remark on below.

The qualified defense of refute-or-accept tacitly opposes various alternatives—norms weaker or more contextually sensitive than refute-or-accept, as well as the rejection of any such norm altogether. The no-norm objection holds that your desire for relevant information and especially the compulsion of reason itself is sufficient to motivate acceptance, if you cannot refute.

The response to these alternatives is dispersed (sections II, III, VII, VIII). My immediate response to the no-norm objection is that it fails to deal squarely with the RAB-dilemma. In a RAB-dilemma, reason does judge the argument favorably. If the violation is of reason for not accepting, the objection dismisses the basic dilemma that animates this paper: There can be warrant to resist reason's judgment that the argument succeeds, despite an inability to refute.

The resolution of the RAB-dilemma in sections VII and VIII is that one can be justified in resisting one's judgment of the cogency of an argument, yet when one does, an epistemic duty is violated. The opposed pulls of the RAB-dilemma are compatible and individually acceptable, so that it is not a true dilemma, though a lingering tension—epistemic and first-personal—remains. Although the resolution depends on the availability of more relevant information to belief than to consciousness or to argumentation, this difference does not explain either the rationality for discomfort with belief's resistance or how argumentation can be a basic epistemic practice, despite restrictions on information, or how the dilemma can arise in examining an argument for oneself, although I concentrate on the dialogue setting.

For developing and responding to the RAB-dilemma, it is useful to focus on refute-or-accept. Though the Lockean objection to it is decisive and the dilemma stands without refute-or-accept, that putative norm should not be simply dismissed. Belief in it, as well as attempts to conform to it, are pressures toward fallacious or distorted reasoning.[3] Positively, a much weaker version of refute-or-accept reflects epistemic demands of argumentation. The resolution of the RAB-dilemma that I propose depends on our limited appreciation of the reasoned basis of our relevant beliefs and related resources (for example, deference to others), as well as restrictions on refutations due to argumentation norms. So the resistance can, I conclude, itself be based on reason—endorsing the steps in an argument is not sufficient for rational persuasion. Even so, a participant will have violated social-epistemic norms of argumentation, since he does not follow where the argument leads.

II. THE STRONG DEMANDS OF REFUTE-OR-ACCEPT

The most forceful implication of refute-or-accept is to exclude a logically unexcludable "middle"—suspension of judgment. The strong demand is manifest in the contrast between the interlocutor, to whom the argument is directed and who is actively engaged with it, and someone who overhears it, though it is not directed to him. An *overhearer*, we will imagine, shares both the interlocutor's interest in the speaker's conclusion and his relevant beliefs. But since the overhearer is not a participant in the argument exchange, refute-or-accept's application to him is attenuated.

If the argument is discovered to fail badly, the interlocutor can blame the arguer himself in a way that is barred from the overhearer, just as an overhearer to a promise cannot blame the false promisor in the direct, personal way that the promisee can. Refute-or-accept forces an explicitness or articulation for the interlocutor's relevant reasons, which is central to argumentation. If the interlocutor does not accept the conclusion, he is bound to produce reasons to fault the argument, unlike the overhearer. Yet what the overhearer has before him is the same premises and inferences as the interlocutor, which alone is what accepting an argument purports to depend on. By virtue of participating in argumentation, the interlocutor incurs the refute-or-accept duty, even if unintentionally, thereby limiting his access to the safe terrain of undecided.

Refute-or-accept is defensible as an instance of reason's requirement to follow—believe—where the argument or evidence leads, once questions of clarification are settled. So acceptance cannot be merely acceptance as a hypothesis or as a supposition for the sake of argument, since the presentation of an argument represents its conclusion as following from true premises. (Here marks a break with the procedural *burden of proof* rules of dialogue models and of legal decision-making: for example, if the prosecution fails to make out its case, the defendant is found not guilty).[4]

---

[3] See Jonathan E. Adler, "Distortion and Argumentation," *Episteme*, IV (2008): 382–401.

However, acceptance as true, which issues in belief, does not require an individual's voluntary control.[5] Dennett does not decide that he does not believe the argument. He finds himself unable to believe it. If, as I assume, we cannot conceive overcoming this inability without applying an indirect ("Pascalian") strategy, the barrier appears to be conceptual.

The exclusion is compatible, though, with the "'ought' implies 'can'" principle. When we challenge (blame) someone for disrespecting the force of argument, the challenge (blame) is directed to belief-related, but voluntary, activities. The interlocutor has neglected objections, or he has failed to shape his dispositions or methods to be fair and impartial in the admission and the evaluation of criticism.

Like the strong notion of acceptance, the "refute" end of refute-or-accept is also very demanding. Just as to accept issues in belief, refutation issues in disbelief (that the argument is cogent). To refute is not merely to criticize (find some flaw, however minor, in) the argument, but to establish its lack of cogency. The ordinary citizen confronted with the libertarian argument is not wholly tongue tied ("But if we raise no public funds, how will we build roads or schools?"). Yet, as he recognizes, he offers no serious opposition to the libertarian's move from basic rights (to the contractual fruits of one's labor) to nondefense or nonpolice related taxation without individual consent as coercive.

Requirements for a refutation are also not met by objections such as *ad hominems*, which include accusations that the arguer's presented reasons are just rationalizations or otherwise ill motivated.[6] The dissent or disagreement of others does not constitute refutation, as appeals to popularity do not justify acceptance. Nor can a refutation be a meta-argument about previously discovered failings of similar kinds of arguments or that the argument presented appears too simple for the issues involved. A poor track record for a category of arguments (for example, proofs of the existence of God) does not refute a new argument within that category. I will refer to these as *external* objections. The justification for excluding these objections bars, as well, rejection on the ground either that norms restrict one's appeal to external objections or that one's failure to refute may only be a weakness of one's own, as the Lockean objection states.

---

[4] In his "Reasoning and Conversation" (*Psychological Review*, CIV (1998): 411–41), Lance Rips tests a number of principles akin to refute-or-accept or burden of proof. See especially his Justification Principles and the Skipped-Response Principle, p. 420. The outcomes provide only equivocal support for these principles. I suspect that the support would have been greater with opinions that involved more personal commitment or investment. Rips also cites a number of formal models of dialogue which maintain that failures "to respond immediately to a claim is a concession," p. 421. See also Frans H. van Eemeren and Rob Grootendorst, *Speech Acts in Argumentative Discussions* (Dordrecht: Foris, 1984), Rule 17b p. 174.

[5] Bernard Williams, "Deciding to Believe," in his *Problems of the Self* (New York: Cambridge, 1973), pp. 136–51.

[6] See further Adler, "Confidence in Argument," *Canadian Journal of Philosophy*, XXXVI (2006): 225–58. For a contrasting view of *ad hominems*, see Alvin I. Goldman, *Knowledge in a Social World* (New York: Oxford, 1999), pp. 150–54.

Another class of restrictions is of deference to one's future self or, more standardly, to others, whom the interlocutor takes to be better positioned to answer such arguments. Since restrictions on deference share a good deal of the basis for the exclusion of external objections, I lump these in as external objections, though distinctive worries apply. Casting doubt on the veracity of a source can constitute a refutation. Still, those sources whose veracity are normally taken for granted acquire a justificatory burden when the speaker's argument renders appeal to them non-neutral or question-begging.

What is the basis for these exclusions or restrictions or imposition of burdens of proof, which apply to external objections, even when they are reliably grounded (as indicative that an argument fails)? Aside from the rationale set out below, they ring true. In a confrontation with his parents' argument that he should stop smoking marijuana, a teenager scores points when he lands the (circumstantial) ad hominen that his parents smoked as well. Yet, the teenager knows that he has entered an argument-stopper, not a refutation, and the parents respond similarly.

The restrictions attempt to purify argumentation, rendering it closer in content to that of professional forums. First, external objections leave open the possibility of knowing that the argument succeeds (that it is cogent). Consequently, second, like the option of suspension of judgment, their admissibility would encourage more opinions to be held partially or to a degree, rather than as definite all-out stands, crucial for engaging argumentation. Third, these criticisms remove the arguer's opportunity to reply by way of elaboration or extension on his argument. Fourth, external objections are not specific to the topic at hand. They distract focus. Fifth, these criticisms hold independent of the content of the argument. Their use is contrary to the interlocutor's representation of himself as open to the argument and to his competence or authority to evaluate it. (Openness does not imply the epistemic possibility that the argument is cogent, which would be sufficient to undermine an interlocutor's contrary belief.) External objections are not responses directed to the content of the arguer's reasons. Sixth, these criticisms deny reciprocal gratitude to the arguer in accord with the practice, which represents the speaker's objective as acceptance within the current exchange.

These six reasons do not apply equally as restrictions on external objections. In particular, objections by deference mainly confront only the third and sixth reasons, and the risk of question-begging. A broader concern with these reasons is their force. Except for the first, which is the only one that is cleanly epistemic, you might very well respond to any of them dismissively in an actual case. Against a defender of a proposed plan to drill for oil in a nature preserve, whom you pointedly remark to be an executive of an oil company, you admit: "I grant that my innuendo that the speaker is prejudiced is based only on his affiliation and that by introducing it I am 'poisoning the wells' for discussion. So what? It is a reliable basis for thinking his argument is likely distorted, and what is at issue is whether we should accept it."

A partial reply is that the fundamental issue is whether argumentation is an epistemic practice. For that issue, the details of my list of external objections and the rationale for their exclusion only need to be broadly correct. The issue is joined by the implication that argumentation restricts the class of acceptable objections, including those reliably based.

The substantive reply is that the harm grounding the exclusions is primarily to the practice of argumentation, not to particular exchanges. The right question to ask is a generalization one: Could the practice regularly realize its objective if it were known to be open to participants to offer external objections freely, where their impact is not blunted by the happenstance of, say, limited memory, interest, or publicity? To this generalization question the six reasons just stated argue for a negative answer. Take deference to the results of an interlocutor's further (future) inquiries. Since complex arguments typically depend on assumptions acceptable only on the speaker's word, there is always latitude for an undemanding: "Seems reasonable. I will investigate." The interlocutor though appends no specific evidence that, for example, a scientific report, cited favorably by the arguer, is flawed or is opposed by an equally respected source. Knowing of this all-purpose resource, arguers would become much less willing to offer their arguments, despite content-appropriate circumstances. Prospects for regular and fruitful engagements between arguer and interlocutor would diminish in turn.[7]

The rationale for these restrictions as keeping argumentation focused on content, facilitating participation in the practice, and preserving argument as yielding proof or knowledge, supports refute-or-accept as well, particularly in the burden it places on suspension of judgment. But this support for refute-or-accept also advances the

---

[7] The restrictions serve the value of articulating intellectual differences and engaging in discussion in order to promote agreement on the account of Allan Gibbard, *Wise Choice, Apt Feelings* (Cambridge: Harvard, 1990), pp. 217–18.

Lockean objection. Since refutation is that much more of an achievement, one's failure to refute is less to the credit of the argument.

III. TESTIMONY AND WEAKENING THE LOCKEAN OBJECTION

In this section, I will attempt to weaken the Lockean objection by drawing a parallel with testimony, as ordinary assertion whose fundamental purpose is to inform.

The parallel rests not only on argument as a form of testimony, involving overt intentional communicative acts. Each aims to transmit or to establish a true proposition. I assume that ordinary testimony is governed by a *default* norm, a version of which is: Accept the speaker's assertion, unless you have specific reason to object. A speaker who offered a hearer directions, and then discovered that the hearer did not follow them or asked another, would be offended. The hearer would shrink before the speaker's "Why did you ask me?" or "Don't you trust me?" Hearers are not to seek evidence of the speaker's trustworthiness, as a condition on their acceptance of his assertion. They too are under a norm that excludes a logically unexcludable "middle."

There are extensive defenses of default norms for testimony elsewhere, but it remains controversial—Is it needed and is it epistemic? To the extent that I draw upon this and other controversial claims from the literature on testimony, my argument is conditional.

Both default rules serve our need to economize and to maintain the cooperative practice. But the economizing does not thereby compromise the standards for an empirical inquiry. In any inquiry, one can always obtain more evidence, reaching greater levels of assurance. Yet, we do end inquiries legitimately. Otherwise we would never reach answers to questions that motivated the inquiry and there would be very few other inquiries that we could undertake. Our finiteness, though, is not any (first-order) reason pro or con a hypothesis, so that the acceptance of that hypothesis is not relativized to limited creatures.

One epistemic line of defense for the default rule is from the overwhelming success of the testimonial practice in transferring good information. Endorsement of argumentation would be implausible were it not similarly, even if not as fully, to yield by-and-large correct judgments, given how important it is to us that our beliefs are true, especially as guiding action.

The default rule for testimony and refute-or-accept matches the speaker's (arguer's) objective in assertion. Speakers enter claims to knowledge; and arguers purport to establish their conclusions. Acceptance is the reciprocal of the speaker's objective. If you have no specific reason to deny that speakers have met their objective, you are bound, within the practice, to fulfill your role (to accept).

As cooperative practices, mutual rules and expectations are required which must be manifestly rational or necessary, so that they can be picked up, or accommodated to, by participants without instruction. For those (hearers, interlocutors) to whom the information is transferred, it is not usually feasible to check on the credentials of their informants (strangers). Yet, a positive initial stance toward the speaker (arguer) seems needed to facilitate the transfer of information. The testimonial problematic begins with hearers lacking evidence about their informants and speakers normally receiving no direct informational benefits from their assertions, which is to set aside the gratitude of the hearer and special—self-interested—motives to inform.

In this reciprocal relation, hearers (interlocutors) also fulfill their ethical or social duty of *fair play*. The duty arises if one has "knowingly participated in and accepted the benefits of a practice acknowledged to be fair."[8] Such a norm provides dependable expectations of how others will respond, assuming that they accept the norm. At a dinner party, you are presented with a plate of (unappetizing) "Spam." The social norms dictate that you eat it, while putting on a "good face," unless you have a special ground (excuse) to object ("My doctor told me to lay off Spam"). Such norms of gratitude and politeness diminish social friction, and so promote, and, through coordination, help to constitute, worthwhile social practices.

Given these parallels between testimony and argument, we should be suspicious of the Lockean objection, which is easily restated to apply to testimony:

> I grant that I cannot find any specific reason to object to your testimony. But that's no reason to accept yours. That I cannot find any specific reason may be just a limit or fault in me. It has nothing to do with whether yours is correct.

This application of the Lockean objection fails to take account of our need for information and our limited ability to acquire it on our own. The objection treats belief's aim to avoid falsity as dominant over the one to seek truth, which involves risk. If the testimonial practice is reliable at least in obviously simple cases, my ratio of large numbers of true to false worthwhile beliefs is expected to be far better if I accept testimony in these circumstances than if I resolutely judge for myself. An informant's judgment may be better than my own in a way that I can anticipate and adjust to accordingly. Despite the frequency of our having only limited knowledge of informants (strangers), testimony is a vast source of good information. Since much of the information that we receive from others guides action (for example, in receiving local driving directions), the practice is constantly subject to testing. Our continued easy and extensive dependence on it is evidence that it regularly passes these tests, so that it is massively, though only tacitly, confirmed.[9]

---

[8] John Rawls, "Justice as Reciprocity," in his *Collected Papers*, Samuel Freeman, ed. (Cambridge: Harvard, 1999), pp. 190–224, p. 210; *A Theory of Justice* (Cambridge: Harvard, 1971), sections 52–53.

IV. ARGUMENT AS SPECIALLY RISKY TESTIMONY

The parallel with testimony weakens or narrows the Lockean objection, but the parallel does not nullify it. In the opening example from Dennett, as well as in the confrontation with the libertarian, I assume that the right response is as presented: a lack of conviction. That response is advanced if argument is an epistemically fragile form of testimony. The interlocutor is much more vulnerable to false persuasion than the hearer to false testimony, pursued in the next section.

The essential differences between argument and testimony is that in argument the arguer extends an invitation to dialogue and *explicitly* presents his reasons. The point of the default-acceptance of testimony is, however, that reasons are not to be presented, unless they are of independent informational value or required to meet a hearer's challenge, subsequent to assertion. Speakers are presumed to have good reason to back their assertion, though it is common knowledge that often they do not retain the originating reasons for their beliefs. The default rule for testimony gains the economy of acceptance without the hearer's explicit evaluation. However, such evaluation is the nature of argumentation, which is the arguer's presumed purpose in presenting his reasons or evidence. A speaker who passes along false testimony can try to excuse himself by noting that "Bill told me," but that excuse will not work for "Bill convinced me."

The presentation, or indication, of reasons is demanded if it is mutual that the conditions for the default-acceptance of testimony are not met. What conditions?

First, and primary, what is argued for are conclusions that are antecedently more tendentious than for ordinary testimony. The speaker should supply reasons to back his assertions because, prominently, what is testified to is either controversial or not within a speaker's presumed domain of competence or lacking in prior plausibility. You will not default accept a speaker's assertion that "Paris Hilton will attend the department's next colloquium."

Second, arguers often have manifestly self-interested reasons to persuade. In ordinary testimony, it is atypical for there to be an incentive to lie or to deceive. You ask a stranger for directions. You believe his response not just due to the evident presumption of his competence, extending to his knowing when he does not know. There is no reason (motive) for him to lie. Important classes of argument, however, particularly on opinionated matters, are noticeably driven by personal, political, and related investments. These generate self-interested reasons to persuade, additional to the pleasure of convincing others or of winning an argument.[10]

---

[9] Adler, *Belief's Own Ethics* (Cambridge: MIT, 2005), chapter 6.

Admittedly, there are numerous, very simple, arguments, not even exchanges, where both differences with testimony are minimal. Aside from implicatures, there are commonplace arguments whose communicative intent is only to inform (or to assist), not persuade: "The bathroom is being cleaned. Use the one downstairs." "You should not throw the newspaper away. It can be recycled." "The Yankees are in first place. The Red Sox must have lost their double-header." However, we cannot abstract to an essential core of simple arguments as we can for testimony. We can, of course, pick out arguments that are simple, as we just did. But for testimony, the pure cases are independently specifiable as those backed immediately only by the word of the stranger. That specification guides abstraction to assertions on topics that presuppose speakers' competence and for which speakers have no motive to deceive (for example, local driving directions).

Not so for argument. The arguer does not invite the interlocutor to accept the argument on just his word, but upon the interlocutor's evaluation. Given explicitness, the pure cases are not those backed by the word of the arguer alone. A highly complex argument can be as paradigmatic as a simple one, since the degree of complexity is content-based, a function of the demands of the conclusion to be reached, the available reasons (evidence), and the relevant beliefs of the audience. In brief, there is no principled epistemic distinction between simple and complex arguments; and the very simple cases, like the above, are not arguments for which refute-or-accept is needed. Since a high percentage of arguments are complex, whereas core cases of testimony are a simple single assertion, arguments are much more susceptible to error in formulation or evaluation than testimony.

The greater susceptibility to error, as well as the previous disanalogies, reveal testimony as closer in its basic epistemics to perception than to argument. The acceptance of perceptual and testimonial reports achieves warrant without an evaluative judgment, unlike for argument. Since arguers explicitly present their reasons, there is no justificatory gap for the reliability of the process to fill. (The exception is for the acceptance of those premises that amount to the acceptance of testimony.) The arguer's stance of explicitness renders pointless the interlocutor's inferring that the best explanation of why the arguer asserted his conclusion is that he believes it for good reason, as is typical for perceptual reports and testimony. For the reasons that the arguer believes the conclusion are right before the interlocutor. The explicitness obviates the need to infer to their (causal) source, unless they are way off.

---

[10] The contrast is weakened during evolutionary times, when much ordinary information is strategically valuable, if Dan Sperber's reasoning is correct in "An Evolutionary Perspective on Testimony and Argumentation," *Philosophical Topics*, xxix (2001): 401–11.

Perception and testimony are plausibly held to "transmit warrant" according to the default rule stated earlier, which I vary for purposes of brevity and comparability:

> If it appears to *X* by perception that *p* in normal (optimal) conditions, then *X* is warranted in believing that *p*, unless *X* has special reason to object.
>
> If *S* asserts that *p* to *H* in normal (optimal) conditions, then *H* is warranted in believing that *p*.[11]

These principles presuppose normalcy conditions in which one judges that *p* only if *p*, and various errors are just interference or noise or outside the domain of optimal functioning, for example, judging colors in harsh light. "Warrant" is partly an objective condition for something like a reliable, evidential, or subjective connection between the processes governing the appearances or activities and the proposition believed. But there are no parallel normalcy conditions for argument such that

> If *X* judges that argument *A* is cogent in normal (optimal) conditions, then *X* is warranted in believing that *A* is cogent.

Consequently, there cannot be a warrant principle for argument akin to those above for testimony or perception. Even under optimal conditions, our reasoning is susceptible to illusions and fallacies. (Think of the fallacy Saul Kripke exposed from 'it was a discovery that Hesperus is Phosphorus' to 'it is contingent that Hesperus is Phosphorus.'[12]) The abstraction that would rule out this possibility would be superhuman reasoning powers and omniscience (to exclude all false assumptions). Omniscience, however, renders the provision of reasons to back an assertion pointless.

---

[11] For entitlement/a priori versions, see Tyler Burge, "Content Preservation," *The Philosophical Review*, CII (1993): 457–88, and Christopher Peacocke, *The Realm of Reason* (New York: Oxford, 2004), chapter 3 and pp. 110–12.

[12] Kripke, *Naming and Necessity* (Cambridge: Harvard, 1972).

V. IMPLICITNESS AND THE VULNERABILITY OF THE INTERLOCUTOR

Although ordinary arguments, unlike testimony, involve the explicit presentation of reasons, the actual explicitness is extremely limited. The Warren Commission's opinionated report on the assassination of President Kennedy is not within our purview.

In everyday argument, arguers need to cite only enough of their reasons so that, given the beliefs or assumptions shared in that context, interlocutors can comprehend how they are meant to establish their conclusions. But that only requires articulating a small portion of the reasons for any serious conclusion. An interlocutor can accept a premise without appreciating (at that time) many of its substantive implications or consequences.[13] There are distortions induced by unrecognized and problematic assumptions, severe practical constraints on interest and time, and limits of imagination. A familiar example is Peter Singer's main principle in "Famine, Affluence, and Morality," which students readily accept. But they are taken aback when led to draw out its apparent implications for their own duties to children suffering from readily curable diseases in distant lands.

Ordinary argument presentations are far less explicit than argument-reconstructions, where premises and conclusion are standardized and which purport to state all assumptions. Much of ordinary argument is implicit, rhetorical (figurative, analogical, allegorical, by example), enthymematic, and by innuendo.[14] It has to be for purposes of brevity, at the least. In actual discourse, prolixity of reasons would yield misunderstanding (an added reason taken as lessening the force of previous ones), rather than greater conviction, unless contrary indicators are present.

Rather than state assumptions or conclusions, the arguer tries to stimulate interlocutors to bridge gaps with shared assumptions that the interlocutor will view as his own, increasing prospects for economic comprehension, as well as for persuasion. The assumptions are not in the interlocutor's focal attention, although tacitly he represents himself as endorsing them. Of course, if an interlocutor finds a gap in an argument, he can challenge the arguer to produce the needed assumptions. But there cannot be any default demand that the arguer produce his assumptions. For the most obvious difficulty, arguers sometimes do not know what assumptions their argument requires.

---

[13] For valuable discussion, see Robert Stalnaker, *Inquiry* (Montgomery, VT: Bradford, 1984), chapter 5.

[14] On implicitness in arguments, see Adler, "Fallacies and Alternative Interpretations," *Australasian Journal of Philosophy*, LXXIV (1994): 271–82, and Larry Wright, "Argument and Deliberation: A Plea for Understanding," this JOURNAL, XCII, 11 (November 1995): 565–85. The prevalence of tacit epistemic processes and relations is explained in David Henderson and Terrence Horgan, "Iceberg Epistemology," *Philosophy and Phenomenological Research*, LXI (2000): 497–535.

Here are just two examples of the large degree of implicitness of everyday argument each standing for a huge number of others. First, arguments where the inference is facilitated by conventional implicatures:

> Tony is 160 pounds. Nevertheless, he should make the linebacker position.

The "nevertheless" acknowledges the interlocutor's presumed doubt that Tony is too light to play linebacker. By the acknowledgment alone, the arguer tries to allay the interlocutor's concern, which facilitates his acceptance of the conclusion.

The second example is from a television commentary. The commentator began with the observation of the liberal-conservative polarization of United States citizens, which has persisted past the 2004 Presidential election. He explains:

> Since both Presidential candidates concentrated on "battleground" states neither candidate directly addressed those holding strongly opposed views. So they surrendered the opportunity to generate dialogue. Thus, polarization was not challenged, and, in fact, with the results of the election, it only became worse.

We easily grasp the commentator's reasoning. Still, it depends on a host of assumptions, notably the following:

> Were the candidates to have confronted the polarized citizens in nonbattleground states, that would have encouraged those citizens to moderate their positions and political attitudes.

Once stated, suspicions are aroused. Yet the argument as presented, along with shared assumptions, purports completeness.

Compatible with our authority to evaluate an argument on a nonspecialized topic of interest, there is much that is relevant, which is hidden from an interlocutor. Now you might think that an easy way out is just for the interlocutor to renounce his authority. But the authority involved is not expertise. To accept a speaker's word is to ascribe to him authority. You will treat his asserting that *p* within bounds of his authority as a content-independent reason to accept it.

We form beliefs and opinions about many subjects with little study or specific knowledge. In order to be active participants in social (conversational) and political life, we must have and defend opinions on a wide range of topics for which it is evident that few have extensive knowledge (from sports to foreign policy to judgments of taste to gossip to...). In Deanna Kuhn's pioneering empirical study,[15] she examines argumentation among ordinary citizens about three topics: "What causes prisoners to return to crime after they're released?" "What causes children to fail in school?" and "What causes unemployment?" (*op. cit.*, p. 16). Despite these questions eliciting a large range of possible causal explanations from subjects, many of which blatantly call for research, she observes

> All but a few participants quite readily offer their views on the topics they are asked about and exhibit little difficulty in expressing themselves (*op. cit.*, pp. 21–22).

In accord with our second reason for restrictions on external objections, these definite positions are held without epistemic qualification (for example, "pretty sure"), despite their invitation to research and expertise.

In this section, argumentation has been shown to place a heavy burden on the interlocutor's judgment. By contrast, the burden in testimony is on the hearer only near the margins (of plausibility), resting otherwise mainly on the reliability of the testimonial process or practice and its constraints on speakers for trustworthy testimony. So, if an argument is seriously engaged, the interlocutor is greatly dependent on trust in the arguer. Many of the arguer's premises, like those alluding to supportive data, are to be accepted on his word alone. Even if an arguer intends nothing devious, he has available a wide variety of devices of persuasion, as well as knowledge of common weaknesses in reasoning that far outstrip our gullibility with core testimony. Consequently, we are much more inclined toward accepting the testimony of strangers than their arguments.

Were it not for the great implicitness of argument, this would be an incongruous result. Given that reasons are provided for arguments, but not for testimony, the degree of vulnerability should be just the reverse. In simple cases, it is. Since argument, unlike testimony, routinely demands evaluation, in simple cases we are less vulnerable to false persuasion with argument than with testimony. But the expectation does not hold up with complex arguments because of implicitness, as well as restrictions on the interlocutor's resources for refutation, especially of external objections. Given that the arguer has presented his reasons, the burden of judgment is with the interlocutor, who cannot appeal to his vulnerability to overrule refute-or-accept without withdrawing from the practice.

---

[15] Kuhn, *The Skills of Argument* (New York: Cambridge, 1991).

VI. REDUCTIOS

The implicitness of argument I raised as a problem for the endorsement of refute-or-accept, since it places a heavy burden on the interlocutor's judgment. But the interlocutor has another defensive resource to which I have alluded. Reductios can refute an argument without locating specific faults and so without requiring appreciation of the argument's implicit assumptions or inferences.

However, this resource is weaker than it appears. Refute-or-accept itself seems implicated in the assimilation of the failure to accept an argument (without finding a specific defect) to reductios. Not to accept an argument without articulating a fault is construed as a reductio refutation.

Refute-or-accept assumes that if you cannot refute the argument, the verdict is not merely negative. Dennett finds the argument presented to him step-by-step credible. Thus, Dennett's resistance to belief appears as a denial of self-trust or authority. We can easily imagine the arguer expressing his dissatisfaction with Dennett's resistance: "You admit that you cannot refute my argument and that you are competent to judge it. It appears to you cogent, yet you do not accept it. That is just a refusal to be guided by reason. If you cannot refute the argument, why don't you accept it—what's left? Are you not at least admitting that presently my conclusion is more warranted than your ignorance or agnosticism about the issue or, more assuredly, your contrary belief?"

An extension of the Lockean objection can be presented as answers to these last questions: We often know that an argument is defective, even when we do not know why it is defective, via, prominently, reductios. However, the reductio category, so understood, is too broad. In RAB-dilemmas, the resistance is either not a reductio or a very weak kind.

First, *informal* reductios, in which an interlocutor rejects an argument for an unacceptable conclusion, do not provide the guarantee of a *formal* reductio. In a formal reductio, an assumption is introduced (for reductio), from which a contradiction—an inherent defect—is derived. A formal reductio is assumptionless and merely exposed. These features—"derived," "contradiction," "inherent defect"—do not hold for informal reductios.[16]

Second, the attribution of a formal reductio is antecedently less dubious than an informal one, an asymmetry that appears to run contrary to the Principle of Charity,[17] since the worse failing would be thought of as the one harder to ascribe "charitably."

---

[16] The contrast depends on a formal reductio leading to a contradiction, not on implications drawn from that contradiction, which is where nonclassical logics typically jump ship. Still, a number of the issues presented here would be viewed differently in either nonclassical logics, as the editors noted, or Bayesian approaches, as Kotzen suggested (personal communication).

The Principle of Charity is on firmest ground in blocking an attribution of recognized inconsistent beliefs. But the dominant cases will be ones in which the arguer fails to recognize the inconsistency. In those cases, a miss is as good as a mile: if you do not recognize a set of your beliefs as inconsistent, though they are, you do not at all suspect it.

With informal reductios the Principle of Charity question arises: Is it more plausible that the arguer believes that unacceptable conclusion or that you, the interlocutor, have erred in attributing it? In favor of the latter alternative is that the arguer will share enough relevant beliefs with the attributor so that he will recognize the attributor's reasons for finding his conclusion unacceptable, even if he thinks that these reasons are outweighed. A Holocaust denier still shares with you beliefs that provide an understanding of why you reject his conclusion.

The arguer snared by a formal reductio, however, cannot even suspect the problem to come, lest he thereby cease holding those beliefs. But the arguer, who is subject to the informal reductio, will share some of those reductio-casting beliefs. Thus, the attributor has a far more difficult—and so less secure—position from which to cast the informal reductio than the formal one.

Third, informal reductios differ in the security with which the argument's conclusion is rejected, and with which they back a necessary second-order claim.

As already noted, informal reductios oppose an argument—a reasoned defense of a claim—by one's own beliefs. Philosophers joke that one man's modus tollens is another's modus ponens. But whatever are the symmetries that this expression alludes to there is an asymmetry within the normal dialectic presentation: The second premise of the modus tollens expresses (merely) the interlocutor's belief, which the arguer's modus ponens—that is, detailed supportive reasons—is offered to overcome.

Typical informal reductios inspire less confidence than those directed against conclusions that are clearly absurd or falsified by a wealth of evidence or contrary to facts, for example, that I have two hands, which Moore held that he is much more convinced of than any argument to the contrary. Other examples that might be mentioned are beliefs —"Moorean facts"— corresponding to denials of: Lewis's realism about possible worlds, conclusions of Zeno's Paradox or sorites, Berkeley's idealism, eliminative materialism, and, of course, radical scepticism.

---

[17] Donald Davidson, *Inquiries into Truth and Interpretation* (New York: Oxford, 1984).

In stark cases like these, we have so sharp a break from the boundary conditions for argumentation that issues of procedures or fairness to the arguer do not really arise. Given satisfaction of boundary conditions, however, the numerous, more familiar, informal reductios provide reasons to reject a conclusion due to incompatibility (of the conclusion) with specific beliefs that are less secure.

Informal reductios require a strong, second-order judgment, not just that a conclusion is not accepted or should not be accepted, but that it is unacceptable. Since that judgment purports to preempt arguments to establish the conclusion, it implies a second-order judgment that nothing will persuade—lead the interlocutor to believe—otherwise, however crudely it is expressed ("that's wild").

The interlocutor construes the presumed beliefs generating his resistance in RAB-dilemmas as less secure. They only constitute *weak* informal reductios, as with the nonlibertarian and the substitutes that I asked you to imagine. In the opening example, Dennett simply fails to be convinced. He says (and with discomfort) "I don't believe..." not "That's crazy [unacceptable]."

The present inability to accept does not deny that further argument or an alternative line of argument might establish the conclusion. Weak, informal reductios are not taken by the interlocutor to justify a second-order judgment that there are no prospects for reasoned persuasion otherwise. For a (strict) informal reductio, the attitude directed toward the arguer is perceived as arrogance: It is so evident that your argument fails, I need not look further than the conclusion. An interlocutor's weak, informal reductio is offered modestly: Your argument seems right, but the conclusion, or some other feature of the argument, makes me suspicious of my own judgment.[18]

Unsurprisingly, then, a reply by weak informal reductios is subject to the restrictions above (section II) against external objections.

---

[18] A standard use of 'unacceptable' for paradoxes should be distinguished from its use as applied to informal reductios. R.M. Sainsbury understands a paradox to be:

> an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises. Appearances have to deceive, since the acceptable cannot lead by acceptable steps to the unacceptable. So, generally, we have a choice: Either the conclusion is not really unacceptable, or else the starting point, or the reasoning, has some nonobvious flaw (*Paradoxes* (New York: Cambridge, 1988), p. 1).

If "unacceptable" is relative to a corpus of beliefs, and *nothing is hidden* (all the relevant information is explicit) and the agent has not erred in either his understanding of this information or the inferences drawn from it, there is no room for an epistemic-metaphysical gap. Now the agent confronts paradox: both options that logically must be open are foreclosed. As I discussed in section v above, the "nothing is hidden" assumption marks a dividing line between paradoxes and RAB-dilemmas: With realistic argument presentations, much is implicit.

In accord with those reasons for restrictions on external objections, they elicit the accusation by the arguer that he was misled to think that the interlocutor's evaluative judgment will be responsive only to the argument and, as with a jury, not turn on antecedent grounds of doubt.

A weak, informal reductio is an epistemic danger to the interlocutor, who casts it. In the case of both testimony and argument, one's merely holding a contrary belief is normally not a sufficient basis for denying opposed testimony (argument). If I believe that the name of the person approaching me is "Tex," and he introduces himself as "Tony," not only do I cease to hold my belief, I come to hold his instead. Similarly, if, on distant memory, I recall that the next train to Penn Station is at 8:10, and my neighbor argues that the arrival is not until 8:50 because of a change in schedule, I also withdraw my belief and, normally, come to hers. Thus, argument and testimony can be prime ways to correct one's belief.

A minimal argument just draws out consequences from your own beliefs. Those consequences may strike you as dubious (an informal self-reductio), undermining either the derivation or the conclusion (belief).[19] In its more typical and salient role as an exchange, the corrective potential of argument is more evident: You do not control the judgments of those with whom you argue or disagree. Like perception, though not as sharply, testimony and argumentation are operationally independent from the beliefs that they can subject to correction. These are processes for the routine acquisition of beliefs, which correct other beliefs as a by-product of their ongoing operation. Argument can only serve as a mechanism of *self-correction*, however, if those beliefs potentially subject to challenge do not determine whether contrary conclusions are correct. Resistance in RAB-dilemmas, as by weak, informal reductios, are in tension with this self-corrective force of argument. In the closing sections I consider whether this tension is really an out-and-out conflict.

VII. REPRESENTATION AND RECONCILIATION

Sections II–VI clarified the support for refute-and-accept and the possibility of legitimate violations of it, while modifying both of these horns of the RAB-dilemma. The overall conclusion dissents from the two evident moves to judge each horn unacceptable. Something like refute-or-accept is viable, despite the Lockean objection. The resistance horn of the dilemma can be justified, despite the interlocutor's failing to follow where the argument leads. Here is a summary of specific findings by section:

---

[19] See Gilbert Harman's contrast (*Change in View* (Cambridge: MIT, 1986), chapter 1) between reasoning and argument (or implication).

II. The rationale for refute-or-accept is to meet the ideal of following where the argument leads, and then the resistance horn is a failure of the interlocutor's commitment to that rationale. The commitment is a very demanding one. It implies a burden of special justification on the logically open option of suspension of judgment. Acceptance issues in belief. External objections do not constitute refutation, even if reliable.

III. Refute-or-accept inherits support from an analogous norm of its respected epistemic parent, testimony, which also exposes a limit on the Lockean objection.

IV. and V. Argument cannot aspire to anything like the reliability of testimony and for this reason, as well as because of the large extent of implicitness of argument, the interlocutor is highly vulnerable to false persuasion.

VI. Weak, informal reductios, based on the resistance horn of the RAB-dilemma, do not reach the security of either formal or informal reductios. The interlocutor does not view the beliefs that back the resistance as justifying claims of the argument-indifferent unacceptability of a conclusion. Weak, informal reductios strain against the self-corrective force of argument, and there are restrictions against them as external objections.

Because these sections offered qualified defenses of both horns of the RAB-dilemma, the task is to reconcile them. The reconciliation returns us to two issues: warrant for resistance in RAB-dilemma and whether refute-or-accept, even if a social norm or a norm of argumentation, is an epistemic one. My positive answer to the latter question explains the possibility of the former.

A formal reconciliation challenges the natural reading of refute-or-accept as claiming that if you recognize that you cannot refute an argument, *you judge properly only if you do accept (believe).* But is this reading correct that refute-or-accept fails if we can rightly not accept, when judging ourselves unable to refute? The reconciliation uncovers an indeterminacy in 'properly' as to whether it refers to a normative standard of argumentation practice or for the unrestricted epistemics of belief itself.

A forceful thesis in epistemology is that in asserting that *p* the speaker *represents himself* as knowing it, a notion that I have already invoked.[20] But the thesis should be interpreted as specifying how the speaker is represented by his intentional performance of the speech-act, rather than by any specific indicator from the speaker himself. If a speaker's assertion represents him as knowing that *p*, then, to develop a claim in section III, a hearer's reciprocal representation is, effectively, the default rule for testimony: The hearer will accept the assertion, unless he has specific reason to object. If the hearer has no specific reason to object, he grants the speaker's representation that he knows. But then the hearer has granted that he has thereby come to know and so represents himself as accepting.

---

[20] The formulation is that of Peter Unger, *Ignorance* (New York: Oxford, 1975), pp. 256–60. The most sustained defense of a variant thesis is that of Timothy Williamson, *Knowledge and Its Limits* (New York: Oxford, 2000), chapter 11.

The assertion-rule-to-default-rule link derives from the core purpose of assertional practice to transmit information. The default rule implies a duty (or commitment) of the hearer's that responds to the earlier noted concern that in testimony the speaker places himself at risk of challenge or blame without prospects of direct informational benefit. If the speaker is bound to back his assertion by knowledge, the hearer represents himself as accepting that assertion, given merely that the hearer does not deny that the speaker has fulfilled his representation. Corresponding to, or reciprocal to, the speaker's representation of himself as knowing is the hearer's representation to accept.[21]

The obvious application is that in engaging in argumentation the interlocutor represents himself as following refute-or-accept. The failure to accept, when the interlocutor is unable to refute, violates the representation, which corresponds to the arguer's claim that the argument is cogent.

---

[21] Compare to Michael Dummett's parallel between the point of assertion as truth and the point of a game as winning. Dummett notes that the player's actual intention does not matter, yet the intention of winning must characterize the game. See Dummett, *Frege: Philosophy of Language* (Cambridge: Harvard, 1981, 2nd edition), pp. 296–302. If the game's point requires such an intention, the proposal from the text is that the opponent must have the reciprocal intention to, for example, defend against an attack, otherwise too the game loses point. On related problems of dialectical models of assertion, see Michael Rescorla, "Assertion and Its Constitutive Norms," *Philosophy and Phenomenological Research* (forthcoming).

Does a converse default-rule-to-assertion-rule link go through by a similar reciprocal representation route? If the hearer is represented as accepting the assertion, the speaker is represented as in the minimal position that guarantees truth, which is the speaker's knowing it. However, what is the "excluded middle" for the speaker that corresponds to that for the hearer? Suggestion: Although there can be no demand to assert corresponding to the hearer's demand to accept, and strictly no need, since acceptance presupposes an assertion, then sincerity excludes strategic retreats to epistemically qualified assertions. Were this option open to speakers, the hearer's typical informational aim will not be satisfied. Some such barrier is defensible from the need for incentives to the speaker due to his informationally uncompensated risk of challenge and blame: if speakers could freely offer epistemically qualified assertions, they would instantly reduce their risk, while still participating (that is, responding usefully and relevantly, if not optimally, to the hearer's informational request).

An interlocutor (or participant, more generally) need not believe that the representation is true of him and he need not even intend to so represent himself. Nevertheless, he remains responsible for satisfying the representation, just as a speaker must respond to challenges to his assertion that *p* of the form 'How do you know that *p*?'. The arguer represents his argument as benefitting the interlocutor, answering his real or presumed inquiry, just as the speaker represents his testimony as for the hearer's benefit. In either case, the presenter already holds the beneficial belief (conclusion, assertion). The representation encourages the arguer to put forth his argument without "losing face," placing a corresponding duty on the interlocutor, comparable to other social norms.

If the interlocutor does not believe accordingly, for, perhaps, Lockean reasons and the failure of default-warrant for argument (end of section v), he is in violation of his reciprocal duty with the arguer. One way an interlocutor's duty is felt is that the interlocutor's confidence in his contrary belief is rationally diminished, as I imagine for the nonlibertarian citizen confronted by the libertarian argument. But that lessening of confidence is not a withdrawal from belief. First, recall that refute-or-accept is not refute-or-withdraw. It is not satisfied by withdrawal without acceptance of the arguer's opposed conclusion, though the withdrawal would amount to a major concession. Second, our confidence in many believed propositions regularly vacillate—specifically diminish—without corresponding loss of belief. A friend's not calling for a long time can lead you to worry as to whether he is a close friend without ceasing to believe it (or retreating to being only pretty sure). If the regular and unremarkable vacillation in confidence with new information were sufficient to erase beliefs, one's corpus of beliefs would be unstable.

What intellectual import does the interlocutor's representation of himself as complying with refute-or-accept have for him, when he violates it in RAB-dilemmas? Among other of its normative consequences, the violation and diminished confidence of the interlocutor provides motivating reasons both to self-criticize the relevant resisting beliefs and to scrutinize the argument and its sources.

The proposal of such a motivating reason may be thought both too weak and too strong. Too weak because it little satisfies the arguer and it seems to disrespect the self-corrective role of argument. The force of this higher-order motivating reason should be judged, however, within the reference class of our vast corpus of beliefs that are mainly taken for granted. Here is a directive to attend to specific beliefs and to attend to them critically, which would otherwise be unexpected, unrequired, and unguided. So it is not empty of force.

But nor is it too strong. The reason to critically examine one's belief does not thereby undermine that belief, as noted three paragraphs back. The nonlibertarian has a motivating reason to reflect critically on his relevant beliefs when he cannot meet the challenge of the libertarian's argument. But his failure does not imply any loss of support for those beliefs, only a rational lessening of his confidence.[22]

VIII. BUT IS IT EPISTEMIC?

Does this reconciliation allow a weakened refute-or-accept to serve as an epistemic norm, rather than at best only a derivative of norms of courtesy, social reciprocity, fair play, and the like? Conversationally presented arguments, though not up to the standards of researched arguments in professional forums, remain an epistemic source of belief, entering unqualified claims of cogency and thereby acceptance.

Why then is not the resistance in RAB-dilemmas obstinacy and a denial of the self-corrective force of argument? The resistance plausibly resides in background beliefs, largely taken for granted, though informing many judgments and actions. If, as is also plausible, these beliefs are taken for granted because evident or well confirmed, the resistance can be warranted. Locating the source of resistance in such beliefs helps to explain why their articulation as a refutation is often unavailable and their force underestimated, as with the nonlibertarian. Those beliefs (reasons) are out of attention and it is pointless to retain beliefs that serve only as support for others. Of these beliefs or others that inform external objections, as with ad hominems or self-doubts about one's own reasoning, the norms that exclude them are readily misconstrued as a denial of their probative force.

The resistance is not a matter of choice or balance. The interlocutor cannot convert his lack of conviction into a refutation, but neither can he just accept the argument. Since belief is nonvoluntary, it is not up to him. Since it is nonvoluntary because of its aim of truth, belief's resistance claims an epistemic source.

The defense of the resistance side of the RAB-dilemma assumes that belief penetrates argument evaluation, rather than entering only as a last stage intervention. Dennett, however, presents his response as an alien force exercising its power over him ("I get to the conclusion, *but I don't believe it!*") with a Moore's Paradoxical ring:

---

[22] On the belief-confidence distinction, see Adler, *Belief's Own Ethics*, *op. cit.*, chapter 10. The distinction is between, roughly, full belief, a threshold reached with the elimination of all possibilities of falsity, and degrees of confidence, which are responsive to a wider set of epistemic reasons. In particular, some external doubts, like learning of disagreement by one's epistemic peers, affect confidence but not full belief. Here I allude to a position that overlaps with, but which is distinct from, the main opposing views. For good examples of these opposed views, see Thomas Kelly, "The Epistemic Significance of Disagreement," in John Hawthorne and Tamar Gendler, eds., *Oxford Studies in Epistemology: Volume 1* (New York: Oxford, 2005), pp. 167–96; and Adam Elga, "Reflection and Disagreement," *Noûs*, XLI (2007): 478–502.

> The argument is convincing, but I don't believe it.

But the alien view seems right only because belief is mistakenly thought of as involved in argument at the end alone, as with informal reductios. But its role is extensive, as the RAB-dilemma and discussion in section II demonstrate.

Corroborative evidence that belief (cognitively) penetrates judgments of cogency is found in the most robust findings from studies on reasoning, which is of *content-effects*. Asked to judge whether premises validly imply a conclusion, subjects are influenced by the prior plausibility that they ascribe (primarily to the conclusion). When all combinations of valid-invalid and believable-unbelievable are tested (on syllogisms), to cite only one strand of research, the findings are that:

> People endorse more valid than invalid conclusions and more believable than unbelievable conclusions. There is also a significant interaction: The belief bias effect is greater for invalid than valid arguments.[23]

Whereas 71% of subjects endorsed invalid-believable syllogisms, only 56% endorse valid and unbelievable ones. More recent work concludes that "there is a general response bias leading to suppression of acceptance of unbelievable conclusions, regardless of the logic of the arguments" (*ibid.*, p. 89).

Objection: How does the appeal to these studies further my ends, since the findings are taken as revealing systematic errors? Reply: The findings show content-dependent influences on judgments of validity, which is evidence of weakness in understanding the distinction between form and content, to put it simply, and intrusive influences on judgments of cogency. However, neither critical implication is contrary to the claim that the content—and so, belief or prior plausibility—dependence is conceptually inevitable and that it yields more accurate judgments of argument acceptability than otherwise (by those of similar cognitive powers, who lacked the same relevant knowledge base). Subjects project from their tasks onto realistic circumstances in which the natural question is one of acceptance, not validity. In this respect, content-effects are a valuable, economical, and unavoidable influence, as well as a self-corrective check, on our judgments, even as they reinforce a poor understanding. Arguments are presented to us as entering truth claims, especially when guides to action, rather than only hypothetically. They are not generated by a random or intentionally difficult mechanism, but by fellow speakers with limited, familiar motives for persuasion, and restraints on their insincerity or unreliability. Consequently, judging implications under the influence of their believability improves, and is expected to improve, our actual judgments.

---

[23] J.S.P. Evans and A. Feeny, "The Role of Prior Belief in Reasoning," in J.P. Leighton and R.J. Sternberg, eds., *The Nature of Reasoning* (New York: Cambridge, 2004), pp. 78–102, p. 88.

Now the resolution of the RAB-dilemma is apparent in a difference between the unbounded and nonvoluntary epistemics of belief and the social or interpersonal epistemics of argumentation. Refute-or-accept imposes restrictions on refutations, though acceptance is not thereby relativized to the norms of argumentation (sections II and VII). However, a failure to accept, while a violation of a duty, follows from the impossibility of forming a belief subject to overt restrictions on truth-relevant reasons. Failure to believe without an articulate reason can nevertheless be truth responsive and respectful of the Lockean objection, even though it violates rational norms of argumentation.

Restrictions on external objections and the heavy demands on the reasoning, knowledge, and intellectual skills of the interlocutor structure argumentation. However, the requirements for argumentation to flourish as an epistemic practice cannot bar some external criticisms and coherence within one's corpus of beliefs from affecting acceptance, since it issues in truth-aiming belief. Given my background beliefs, it is more credible that I wrongly judge that an elegant argument to justify robbing the rich to aid the poor works than that the beliefs that block persuasion are merely stubborn or dogmatic. Philosophers know that it is wrong to beg the question. Nevertheless, there are some questions that should—must—be begged: prominently, denials of Moorean facts. Similarly, libertarian arguments can be formulated so that it begs their question to appeal to various widely respected (liberal) sources to challenge their arguments. Yet that resistance can be warranted.

The resolution defended accords with our reaction to, and within, RAB-dilemmas. It contributes to explaining the entitlement of ordinary citizens to maintain many of their salient beliefs, despite frustrating exchanges not only with libertarians, but also with clever lawyers, smooth politicians, credentialed academics, or other folk who catch them off guard. As befits a dilemma, the resistance by belief, even when warranted, is a resistance under tension. Whereas with rejection (via informal reductios), the interlocutor recognizes, understands, and stands behind his clear, conflicting beliefs, with the nonacceptance of weak, informal reductios, the resistance is that one's relevant beliefs do not budge, which the interlocutor may neither well appreciate nor stand behind.

The strain between the social epistemics governing a participant, who violates a shared norm of the practice, and the individual epistemics, which underlies the violation, often, admittedly, does not give rise to intellectual tension or motivation toward self-critical scrutiny. However, an indifferent reaction loses force as a datum under the generalization question introduced in section II above. Applied here, the generalization question asks what would happen if you were always free to respond indifferently in an RAB-dilemma. In sections II and VII, we noted that argumentation would diminish as a mechanism of self-correction, among other epistemic costs. The numerous occasions when we are indifferent to our violations of refute-or-accept are misleading to the practice functioning optimally under normal, though practically constrained, conditions.

The resulting conflict in first-personal thought is hinted at in the question: How can I continue to self-ascribe authority, while resisting my own judgment of cogency, and how can I distinguish my resistance from dogmatism or intellectual weakness? The testimonial practice does not press this first-personal conundrum when the hearer just does not accept an assertion. While the hearer violates a norm, and one that justifies the frustrated expectation of the speaker, the hearer need not be in conflict with an evaluation of his own. But in an RAB-dilemma, the interlocutor takes himself to fail to follow where the argument leads, as he judges it. The interlocutor thereby imposes normative pressure on himself, a common resolution of which, unlike Dennett's, is credulity toward, or distortion of, another's argument, rather than tolerance of one's intellectual discomfort, perplexity, or ambivalence.

JONATHAN E. ADLER

Brooklyn College and the
Graduate School, City University of New York

CAMBRIDGE

New and Noteworthy Titles

Genetic Analysis
A History of Genetic Thinking
Raphael Falk
Cambridge Studies in Philosophy and Biology
$90 00: Hb: 978-0-521-88418-1:
344 pp.

Nietzsche: *Writings from the Early Notebooks*
Edited by
Raymond Geuss
and Alexander Nehamas
Translated by
Ladislaus Löb
Cambridge Texts in the History of Philosophy
$85 00. Hb. 978-0-521-85584-6
$22.99: Pb: 978-0-521-67180-4:
328 pp.

Kierkegaard: *Concluding Unscientific Postscript*
Edited and translated by
Alastair Hannay
Cambridge Texts in the History of Philosophy
$108.00: Hb: 978-0-521-88247-7
$39.99: Pb: 978-0-521-70910-1:
582 pp.

The Virtue of Aristotle's Ethics
Paula Gottlieb
$85.00· Hb: 978-0-521-76176-5:
262 pp.

The Three Cultures
Natural Sciences, Social Sciences, and the Humanities in the 21st Century
Jerome Kagan
$70.00: Hb: 978-0-521-51842-0
$21.99: Pb: 978-0-521-73230-7:
328 pp.

The Philosophy of Death
Steven Luper
$90.00: Hb: 978-0-521-88249-1
$28.99 Pb: 978-0-521-70912-5.
264 pp.

The Cambridge Companion to Boethius
Edited by John Marenbon
Cambridge Companions to Philosophy
$90.00. Hb· 978-0-521-87266-9
$32.99: Pb: 978-0-521-69425-4.
372 pp.

The Cambridge Companion to Constant
Edited by
Helena Rosenblatt
Cambridge Companions to Philosophy
$90.00: Hb: 978-0-521-85646-1
$29.99: Pb: 978-0-521-67243-6:
448 pp.

Ludic Proof
Greek Mathematics and the Alexandrian Aesthetic
Reviel Netz
$99.00: Hb: 978-0-521-89894-2:
272 pp.

An Introduction to Decision Theory
Martin Peterson
Cambridge Introductions to Philosophy
$80.00: Hb: 978-0-521-88837-0
$26.99: Pb: 978-0-521-71654-3:
328 pp.

The Uses of the Past from Heidegger to Rorty
Doing Philosophy Historically
Robert Piercey
$90.00: Hb: 978-0-521-51753-9.
234 pp.

Kant's *Idea for a Universal History with a Cosmopolitan Aim*
A Critical Guide
Edited by
Amélie Oksenberg Rorty
and James Schmidt
Cambridge Critical Guides
$99.00 Hb 978-0-521-87463-2:
270 pp.

Reasons for Action
Edited by David Sobel
and Steven Wall
$90.00: Hb: 978-0-521-87746-6:
298 pp

How We Get Along
J. David Velleman
$80.00: Hb: 978-0-521-88853-0
$24.99: Pb: 978-0-521-04340-3·
232 pp.

Prices subject to change

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

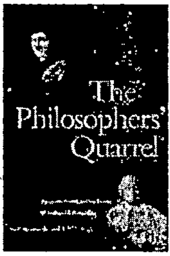
1584 · 2009
425 YEARS OF CAMBRIDGE
PRINTING AND PUBLISHING

PHILOSOPHY from YALE

The Philosophers' Quarrel

Rousseau, Hume, and the Limits of Human Understanding

ROBERT ZARETSKY AND JOHN T. SCOTT

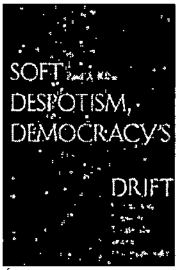


"**An exceptional contribution** to the study of both western political philosophy and the history and politics of 18th-century France. There are fascinating insights on every page." —Roger D. Masters 10 illus $27.50

Soft Despotism, Democracy's Drift

Montesquieu, Rousseau, Tocqueville, and the Modern Prospect

PAUL A. RAHE



"This is an exemplary deployment of great past thinkers in **an intensely provocative, deliberately controversial meditation** on the profound strengths and weaknesses or dangers in our political culture." —Thomas L. Pangle $38.00

Life Explained

MICHEL MORANGE

Translated by Matthew Cobb and Malcolm DeBevoise

"**Morange approaches the question 'what is life'** thoughtfully and with an open mind. He accepts what seems sensible and rejects that which is overblown or confusing." —William C. Summers

An Odile Jacob Book
A Caravan Book: caravanbooks.org $25.00

The Kingdom of Infinite Space

An Encounter with Your Head

RAYMOND TALLIS

This **"highly recommended"** *(Choice)* book blends science and philosophy to explore how the head is connected to our sense of identity and consciousness and the astonishing range of processes and phenomena that go on within it.

$28.00

Long Commentary on the *De Anima* of Aristotle

AVERROES (IBN RUSHD) OF CORDOBA

Translated with Introduction and Notes by Richard C. Taylor
With Thérèse-Anne Druart, Subeditor

This translation of Averroes' Long Commentary on Aristotle's *De Anima* brings to English-language readers **a central work of medieval philosophy.**

Yale Library of Medieval Philosophy Series $85.00

André Masson and the Surrealist Self

CLARK POLING

This richly documented book examines the attempts of the French Surrealist artist André Masson (1896–1987) to define "self" in his art in the period between the early 1920s and 1940, the most fruitful period of classic Surrealism, culminating in the emergence of existentialism.

100 b/w + 33 color illus. $50.00

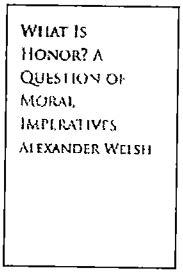
YALE University Press • yalebooks.com

PHILOSOPHY from YALE

What Is Honor?

A Question of Moral Imperatives

ALEXANDER WELSH



What Is Honor? A Question of Moral Imperatives

Alexander Welsh

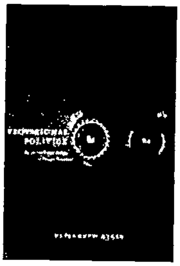
"**Intelligent, well-informed, and thought-provoking,** this book maps and renders intelligible a formidably difficult terrain: the intersection of honor with ethics in both literature and philosophy. This is an important book."—A.D. Nuttall

$38.00

Provisional Politics

Kantian Arguments in Policy Context

ELISABETH ELLIS



"**This is the most original use of Kantian ideas** and insights by a theorist of contemporary politics since John Rawls. . . . It establishes [Ellis] as one of the leaders of the rising generation of political theorists."—Ian Shapiro

$50.00

New in paper

Confucius

A Life of Thought and Politics

ANN-PING CHIN

"**A passionate biography** . . . [and] a fascinating attempt to recover the actual historical figure who did so much to shape one of the world's great civilizations."—Jonathan Lear

$14.95

Babies by Design

The Ethics of Genetic Choice

RONALD M. GREEN

"**In this clear-eyed and generally optimistic book,** both promise and risk are ably weighed and balanced. The science is clearly explained, and there are signposts to help guide us through the moral maze."—*Economist*

A Caravan Book: caravanbooks.org $17.00

Why Arendt Matters

ELISABETH YOUNG-BRUEHL

"Young-Bruehl repeatedly and **successfully unpacks Arendt's views** of such concepts as action, power, forgiveness, judgment, radical evil, revolution, and the human condition itself."—Carlin Romano, *Chronicle of Higher Education* $14.00

The Persistence of Poverty

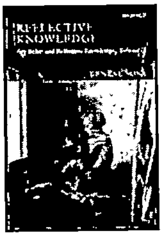
Why the Economics of the Well-Off Can't Help the Poor

CHARLES KARELIS

"This book is **astonishingly lively.** It ranges from the facts of life among the American underclass to deep philosophical puzzles about what does and does not count as 'rational' behavior."—Alan Ryan 6 illus. $18.00

YALE University Press • yalebooks.com

New from OXFORD



Decision Theory and Rationality
JOSÉ LUIS BERMÚDEZ
A challenging new interpretation of one of the key theoretical tools in the human and social sciences.
2009 176 pp. $50.00

Well-Being and Death
BEN BRADLEY
Ben Bradley investigates what is good about life and what is bad about death.
2009 224 pp. $60.00

The Myth of Mars and Venus
Do Men and Women Really Speak Different Languages?
DEBORAH CAMERON
"A wonderfully refreshing new book.... A brilliantly detailed review of where and when different and incompatible stories are made to fit our culture."
—*Feminist Review*
2008 192 pp. $19.95

Perception, Hallucination, and Illusion
WILLIAM FISH
The first full-length treatment of disjunctivism about visual experiences.
(Philosophy of Mind Series)
2009 208 pp. $49.95

Practical Intelligence and the Virtues
DANIEL C. RUSSELL
Daniel Russell develops the idea that a necessary part of virtue is practical intelligence.
2009 416. pp. $99.00

Darwinian Populations and Natural Selection
PETER GODFREY-SMITH
Godfrey-Smith draws on new developments in biology, philosophy of science, and other fields to give a new analysis and extension of Darwin's idea.
2009 224 pp. $49.95

The Development of Modern Logic
Edited by **LEILA HAAPARANTA**
This volume contains newly-commissioned articles covering the development of modern logic from the late medieval period through the end of the twentieth-century.
2009 1008 pp. $150.00

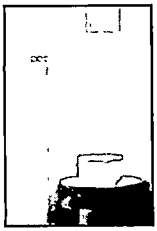
Effective Intentions
The Power of Conscious Will
ALFRED R. MELE
Is free will an illusion? Some scientists say yes; Alfred Mele says no.
2009 192 pp. $45.00

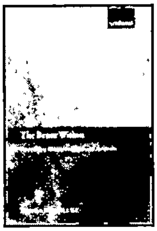
Reflective Knowledge
Apt Belief and Reflective Knowledge, Volume II
ERNEST SOSA
A collection of groundbreaking work in epistemology.
2009 240 pp. $55.00

Visit us at www.oup.com/us to sign up for our philosophy e-newsletter

OXFORD
UNIVERSITY PRESS
To order, or for more information, please call 1-800-451-7556.
Visit our website at www.oup.com/us.

New from OXFORD





NEW IN PAPERBACK

The Possibility of Knowledge
QUASSIM CASSAM
In this highly original book Quassim Cassam explains how epistemological questions arise and how they should be answered.
2009(Hardback 2007) 256 pp.
Paperback $35.00 Hardback $60.00

Justice and the Social Contract
Essays on Rawlsian Political Philosophy
SAMUEL FREEMAN
"Freeman is the leading authority on the thought and writing of John Rawls, and Rawls was the leading political and social philosopher of the 20th Century."
—Thomas Nagel, NYU
2009(Hardback 2006) 352 pp.
Paperback $24.95 Hardback $55.00

Agency and Deontic Logic
JOHN F. HORTY
"A major advance in the field."
—*Notre Dame Philosophical Reviews*
2009 (Hardback 2001)
208 pp. 33 halftones
Paperback $29.95 Hardback $75.00

On Images
Their Structure and Content
JOHN V. KULVICKI
"One of the most valuable and important contributions to the topic."—*Mind*
2009 (Hardback 2006)
272 pp. 25 halftones
Paperback $35.00 Hardback $55.00

The Brute Within
Appetitive Desire in Plato and Aristotle
HENDRIK LORENZ
A comprehensive study of Plato's and Aristotle's conceptions of non-rational desire.
2009 (Hardback 2006) 240 pp.
Paperback $45.00 Hardback $99.00

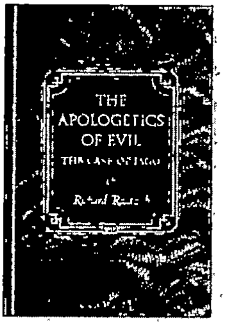
The Oxford Handbook of Bioethics
Edited by BONNIE STEINBOCK
Am authoritative, state-of-the-art guide to current issues in bioethics.
(Oxford Handbooks in Philosophy)
2009 (Hardback 2007) 768 pp Paperback $55.00 Hardback $150.00

Thought's Footing
A Theme in Wittgenstein's Philosophical Investigations
CHARLES TRAVIS
An enquiry into the relationship between the ways things are and the way we think and talk about them.
2009 (Hardback 2006) 240 pp. Paperback $35.00 Hardback $65.00

Natural Moralities
A Defense of Pluralistic Relativism
DAVID B. WONG
"An exemplary study."
—Barbara Herman, UCLA
2009 (Hardback 2006) 312 pp.
Paperback $24.95 Hardback $50.00

OXFORD
UNIVERSITY PRESS

To order, or for more information, please call 1-800-451-7556.
Visit our website at www.oup.com/us.



The Apologetics of Evil

The Case of Iago

Richard Raatzsch

Translated by Ladislaus Löb

"What is evil? What are its forms? How is it motivated? Raatzsch's treatment of [these questions] is sensitive, imaginative, and broadly based. This book brings together different lines of argument from epistemology, moral philosophy, and philosophy of mind in a highly compact and supercharged, yet fully comprehensible form. The result packs an enormous intellectual punch."
—Raymond Geuss, University of Cambridge

Princeton Monographs in Philosophy
Harry G. Frankfurt, Series Editor
Cloth $26.95 978-0-691-13733-9

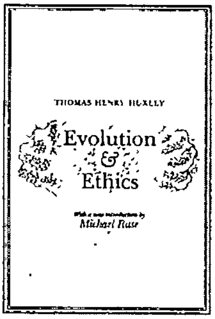
Philosophical Essays, Volume 2

The Philosophical Significance of Language

Scott Soames

The two volumes of *Philosophical Essays* bring together the most important essays written by one of the world's foremost philosophers of language. The essays in Volume 2 illustrate the significance of linguistic concerns for a broad range of philosophical topics—including the relationship between language and thought; the distinction between metaphysical and epistemic possibility; the nature of necessity, actuality, and possible worlds; and skepticism about meaning and mind.

Paper $39.50 978-0-691-13683-7
Cloth $95.00 978-0-691-13682-0



Evolution and Ethics

Thomas Henry Huxley

Evolution and Ethics, widely considered to be Thomas Henry Huxley's greatest lecture, distilled a lifetime's wisdom and sensitive understanding of the nature and needs of humankind. This new edition features an introduction by renowned historian and philosopher of biology Michael Ruse, placing Huxley's lecture in its original context while showing its even deeper relevance for our own time.

Paper $16.95 978-0-691-14130-5
Cloth $39.50 978-0-691-14192-3

PRINCETON UNIVERSITY PRESS press.princeton.edu